NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

DISSERTATION

SUPERQUANTILE REGRESSION: THEORY, ALGORITHMS, AND APPLICATIONS

by

Sofia I. Miranda

December 2014

Dissertation Supervisor: Johannes O. Royset

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 2014	Dissertation (SEP 2011 - DEC 2014)

4. TITLE AND SUBTITLE	5. FUNDING NUMBERS
SUPERQUANTILE REGRESSION: THEORY, ALGORITHMS, AND APPLICATIONS	
6. AUTHOR Sofia I. Miranda	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number N.A.

12a. DISTRIBUTION/AVAILABILITY STATEMENT	12b. DISTRIBUTION CODE
Approved for public release; distribution is unlimited	

13. ABSTRACT(*maximum 200 words*)

We present a novel regression framework centered on a coherent and averse measure of risk, the superquantile risk (also called conditional value-at-risk), which yields more conservatively fitted curves than classical least squares and quantile regressions. In contrast to other generalized regression techniques that approximate conditional superquantiles by various combinations of conditional quantiles, we directly and in perfect analog to classical regression obtain superquantile regression functions as optimal solutions of certain error minimization problems. We show the existence and possible uniqueness of regression functions, discuss the stability of regression functions under perturbations and approximation of the underlying data, and propose an extension of the coefficient of determination R-squared and Cook's distance for assessing the goodness of fit for both quantile and superquantile regression models. We present two classes of computational methods for solving the superquantile regression problem, compare both methods' complexity, and illustrate the methodology in eight numerical examples in the areas of military applications, concerning mission employment of U.S. Navy helicopter pilots and Portuguese Navy submariners, reliability engineering, uncertainty quantification, and financial risk management.

14. SUBJECT TERMS	15. NUMBER OF PAGES	16. PRICE CODE
Superquantile, superquantile regression, buffered reliability, uncertainty quantification, surrogate estimation, superquantile tracking, dualization of risk	147	

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18 298-102

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

SUPERQUANTILE REGRESSION:
THEORY, ALGORITHMS, AND APPLICATIONS

Sofia I. Miranda
Lieutenant, Portuguese Navy
B.S., Naval Military Sciences, Portuguese Naval Academy, 2003
M.S., Operations Research, Naval Postgraduate School, 2011
M.S., Applied Mathematics, Naval Postgraduate School, 2011

Submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN OPERATIONS RESEARCH

from the
NAVAL POSTGRADUATE SCHOOL
December 2014

Author: Sofia I. Miranda

Approved by:
Johannes O. Royset
Associate Professor of
Operations Research
Dissertation Supervisor

Carlos F. Borges
Professor of Applied
Mathematics

Lyn R. Whitaker
Associate Professor of
Operations Research

Javier Salmerón
Associate Professor of
Operations Research

Nedialko B. Dimitrov
Assistant Professor of
Operations Research

Approved by: Robert F. Dell
Chair, Department of Operations Research

Approved by: Douglas Moses
Vice Provost for Academic Affairs

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

We present a novel regression framework centered on a coherent and averse measure of risk, the superquantile risk (also called conditional value-at-risk), which yields more conservatively fitted curves than classical least squares and quantile regressions. In contrast to other generalized regression techniques that approximate conditional superquantiles by various combinations of conditional quantiles, we directly and in perfect analog to classical regression obtain superquantile regression functions as optimal solutions of certain error minimization problems. We show the existence and possible uniqueness of regression functions, discuss the stability of regression functions under perturbations and approximation of the underlying data, and propose an extension of the coefficient of determination R-squared and Cook's distance for assessing the goodness of fit for both quantile and superquantile regression models. We present two classes of computational methods for solving the superquantile regression problem, compare both methods' complexity, and illustrate the methodology in eight numerical examples in the areas of military applications, concerning mission employment of U.S. Navy helicopter pilots and Portuguese Navy submariners, reliability engineering, uncertainty quantification, and financial risk management.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Johannes Royset, for his time, patience, and guidance throughout this arduous but rewarding journey. I am grateful for having the chance to work with you and for the knowledge you transmitted along this process. Even when I returned home to serve the Portuguese Navy, you were tireless. I am so glad you did not give up on me. Muito obrigada!

Also, I also want to express my gratitude to the other members of my committee, Dr. Carlos Borges, Dr. Lyn Whitaker, Dr. Javier Salmerón, and Dr. Ned Dimitrov. I am sincerely grateful to you for sharing your comments, suggestions, and revisions during the completion of this dissertation. Dr. Matt Carlyle and Dr. Ron Fricker, along with the other Operations Research faculty of the Naval Postgraduate School, thank you for allowing me the opportunity to learn from you. I am forever indebted to Dr. R. Tyrrell Rockafellar for his expertise, and Dr. Stan Uryasev for his support with Portfolio Safeguard implementations.

My fellow Ph.D. students, Jay Foraker, Jesse Pietz, Dick McGrath, and Christian Klaus, or shall I say Dr. Jay, Dr. Jesse, Dr. Dick, and Dr. Christian, congrats on your achievements and I am grateful for your friendship and encouragement during some difficult moments. Matt Hawks, Gary Lazzaro, and Austin Wang, just keep up the hard work. You will do great!

This achievement is only possible because the Portuguese Navy believed in me and I will do my very best to show that this level of education is a strategy that should be constantly pursued.

I would like to thank my family and friends for their constant care and support. But most of all, I am grateful to you, my husband, Paulo. You gave me courage, and you showed me (a few times) that giving up is not for me. Your constant support made me strong to endure the ups and downs we experienced during this memorable journey. And I certainly will not forget when you took care of our baby when we

returned to Monterey for almost two months, in July 2014. I am glad you were able to witness all those great baby milestones and I am so proud of you as a Submariner, as a Husband, and now also as a great Dad.

And finally, this dissertation is for you, my baby girl, Isabel. I love having you in my arms, and although you are still too young to understand what a hug is, your warmth has given me the strength and audacity to keep moving forward and complete this goal I have set for myself.

EXECUTIVE SUMMARY

Analysts are often concerned with upper-tail realizations of random variables describing loss, cost, damage of a system and attempt to approximate such loss random variables in terms of explanatory random variables that are more accessible in some sense. We develop a novel regression framework that naturally extends least squares and quantile regressions to contexts where an analyst seeks to assess regression errors not by squaring them, as in the case of least squares regression, or by looking at their signs, as in the case of quantile regression, but by weighing larger errors increasingly heavily in a way consistent with a coherent and averse risk measure, the superquantile risk measure (also called conditional value-at-risk).

In contrast to other generalized regression techniques that approximate conditional superquantiles by various combinations of conditional quantiles, this framework for *superquantile regression* is the first attempt to use superquantiles directly in a regression model. The only assumption we require is that the involved random variables have finite second moment. We rely on the superquantile-based risk quadrangle and use the corresponding relations between measures of deviation, risk, and error applied to the superquantile as the statistic to obtain superquantile regression functions as optimal solutions of an error minimization problem. We develop the fundamental theory for superquantile regression and build an alternative problem, the deviation-based superquantile regression problem, which determines the regression coefficients by minimizing a measure of deviation as opposed to a measure of error, leading to computational advantages in problem size and simplification of the objective function. We examine existence and uniqueness of the obtained regression functions as well as consistency and stability of the regression functions under perturbations due to possible measurement errors and from approximating empirical distributions generated by samples of the underlying data. We develop rate of convergence results under mild assumptions.

In this dissertation, we construct a model validation technique by extending the concept of coefficient of determination used in least squares regression to both quantile and superquantile regression. We show that these coefficients of determination are bounded between 0 and 1, with values near 1 preferred, and we also demonstrate that the superquantile regression problem in fact maximizes the coefficient of determination when it aims to minimize the error of the loss random variable by wisely selecting the regression coefficients. Since adding explanatory random variables possibly increases the coefficient of determination, we define an adjusted coefficient of determination for quantile and superquantile regression. Another validation analysis tool that we develope is the concept of Cook's distance applied to quantile and superquantile regression.

We present two classes of computational methods for solving superquantile regression problems. The first computational method is denoted primal method, where we minimize the superquantile deviation measure using analytical integration or numerical integration schemes. The second computational method is based on the dualization of risk. We build a new superquantile regression problem by using the expression of risk and deviation. We compare the complexity of the methods and demonstrate which ones are more efficient according to the data size and show that dual methods are superior and only marginally slower than methods for least squares regression.

Finally, we present a series of numerical examples that show some of the application of superquantile regression, such as superquantile tracking and surrogate estimation, that we encounter in the areas of financial risk management, military applications, reliability engineering, and uncertainty quantification. We compare computational methods by presenting their runtimes and see how the coefficient of determination and the adjusted one can be relevant in assessing the goodness of fit of the obtained regression models.

I. INTRODUCTION

A. MOTIVATION AND BACKGROUND

One of the major concerns among analysts is how to address random variables describing possible "cost," loss, and "damage," but for which there is incomplete distributional information available. A possibility is to attempt to approximate such a loss random variable by a combination of explanatory random variables that are more accessible in some sense. This situation naturally leads to least squares regression and related models that estimate conditional expectations. While such models are adequate in many situations, they fall short in contexts where a decision maker is risk averse, i.e., is more concerned about upper-tail realizations of the loss random variable than average loss, and views errors asymmetrically with underestimating losses being considered more prejudicial than overestimating.

Another approach is based on quantile regression (see Koenker, 2005; Gilchrist, 2008 and references therein), which accommodates risk-averseness and an asymmetric view of errors by estimating conditional quantiles at a certain probability level such as those in the tail of the conditional distribution of the loss random variable. While suitable in some contexts, quantile regression only deals with the signs of the errors and therefore might be overly "robust" in the sense that portions of a data set can change without necessarily impacting the best-fit regression function, as illustrated below.

In this dissertation, we focus on contexts where a decision maker is concerned about upper-tail realizations of the loss random variable, and errors are not only seen asymmetrically but their magnitude is also taken into account. Of course, a parallel development with an opposite orientation, focused on profits and gains, and concerns about overestimating instead of underestimating is also possible but not considered in this dissertation.

Before we proceed with the literature review, we analyze one simple example.

We consider a loss random variable Y and an available explanatory random variable X. Since the distribution of the loss random variable Y might not be fully known, it may be beneficial to approximate Y by this random vector X.

For this example, we have a table of 50 pairs of observations available, $\{x^i, y^i\}$, with $i = 1, ..., 50$, as seen in the scatter plot in Figure 1. We consider a regression function of the form $f(x) = c_0 + cx$, with c_0, $c \in \mathbb{R}$. This numerical example is artificially designed to show how different regression techniques react to small data changes.

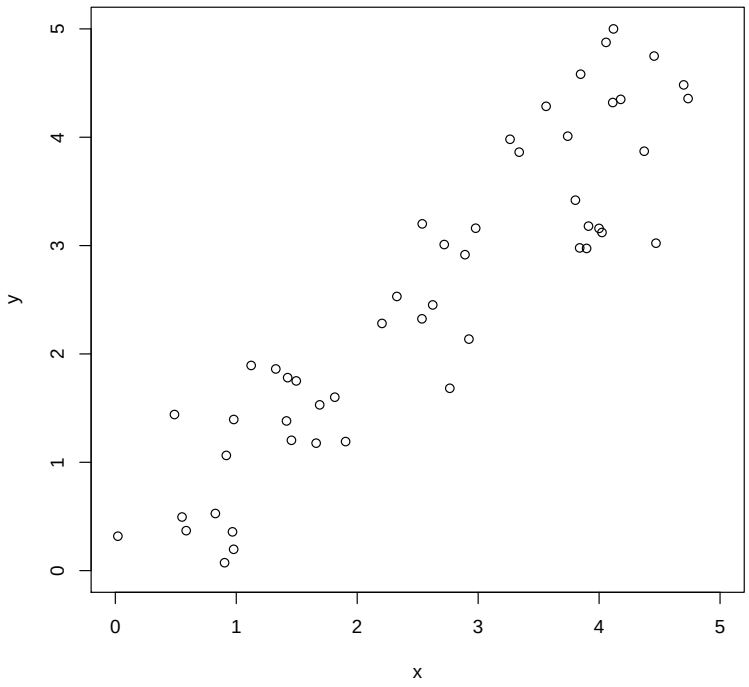
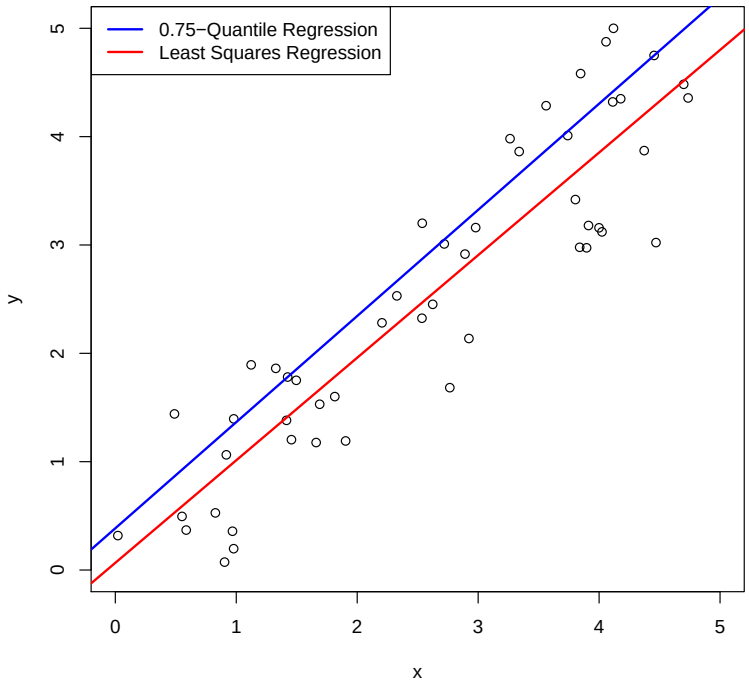
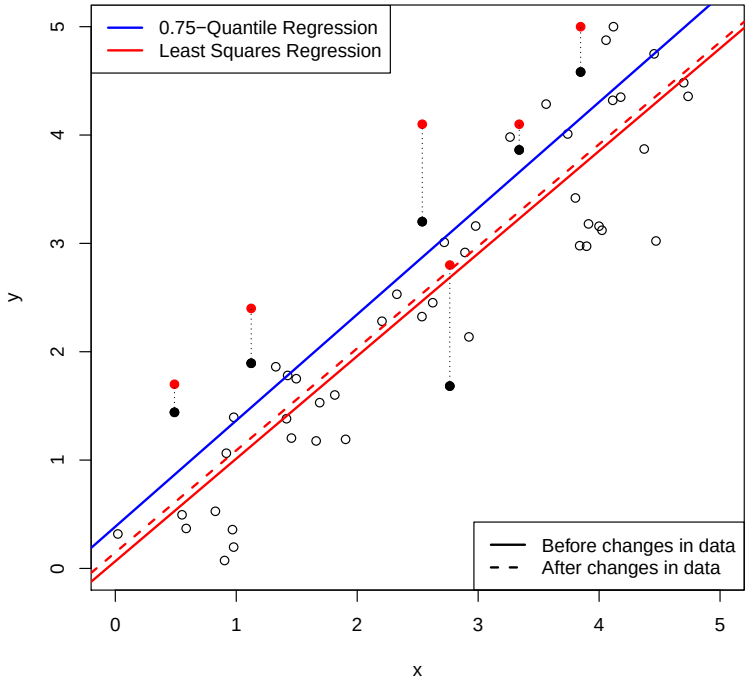


Figure 1. Scatter plot of the data for the constructed example.

Figure 2(a) gives the least squares and 0.75-quantile regression functions. We observe that the 0.75-quantile regression function divides the data set into two, such that 25% of the observations remain above the obtained regression, while the remaining 75% lay below.



(a) Before any changes in the data set.



(b) After shifting six observations upwards.

Figure 2. Least squares regression vs. quantile regression at a probability level $\alpha = 0.75$, before and after some changes in the data set.

In Figure 2(b), we see how the least squares and the quantile regression models adjust to changes in the data set, denoted by the red dots. Notice that the observations are moved upwards without changing their position relative to the 0.75-quantile regression curve. The balance of 25% of the observations above and 75% of them below the quantile regression curve has not been compromised. Therefore, as we can observe in Figure 2(b), the quantile regression curve does not shift after modifying the six observations. Such robustness is sometimes desirable, but at other times there is the need for responsiveness. In comparison, the least squares regression function shifts upwards reacting to the data changes.

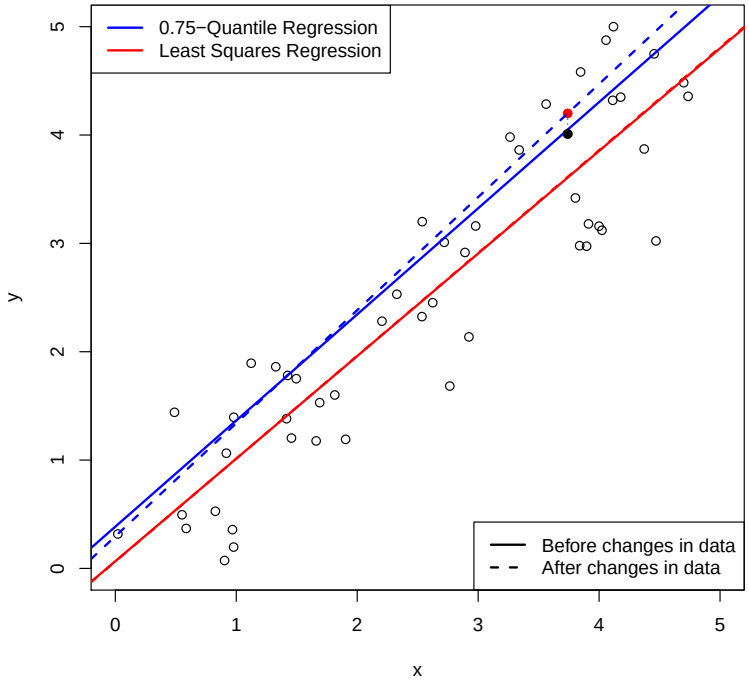


Figure 3. Least squares regression vs. quantile regression at a probability level $\alpha = 0.75$, before and after changing one observation in the data set.

Changing only one observation, as shown in Figure 3, we note that the obtained quantile regression function changes its slope, while the least squares regression function hardly changes. If we shift this observation even further, the change in the

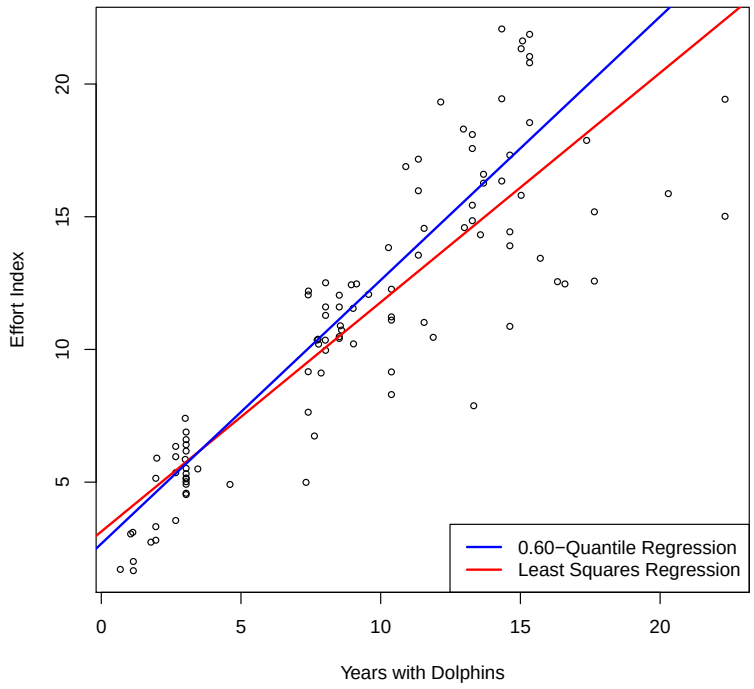


Figure 4. Least squares vs. quantile regression at a probability level $\alpha = 0.60$.

slope for the quantile regression function is even more significant. Once again the least squares regression model hardly changes. If we change this observation in red even further upwards, we would notice no more changes in the quantile regression function obtained in Figure 3, since the balance of the data above and below the quantile regression would no longer be compromised.

Quantile regression is a robust regression technique, but its sensitivity to changes in data might sometimes be too small as indicated above. Other times the sensitivity might be too large as illustrated above where the change of a single data point triggers a jump in the regression curve. On the contrary, the least squares regression is more stable, with smooth adjustments in the curve comparable to changes in the data set.

As another motivation to this novel regression technique, we consider a real-world data set: the Portuguese Navy submariners effort index, provided by the Portuguese Navy Submarine Squadron. In this data set we seek to estimate the random

variable Y that represents the effort index of the submariners. This index was created as a decision tool to support human resource management inside the Submarine Squadron. It allows planners to assess which submariners are "due" for another mission.

In Figure 4, we have 103 observations of number of years since a submariner has gained the insignia of the Portuguese submarine service (X_{dolphins}) against the submariners effort index (Y). In red and blue colors, we see two regression functions, the least squares and the 0.60-quantile regression, respectively. The 0.60-quantile regression fit analyzes the sign of the errors defined as the differences between the loss random variable Y and the chosen linear model. Instead of only regarding the signs of these errors, we want to also account for their magnitudes, namely we want to analyze the average of the 40% highest effort indices.

These two examples motivate the need to move beyond least squares and quantile regression and develop superquantile regression. They illustrate how a regression technique such as the quantile regression, which accommodates risk-averseness and an asymmetric view of errors, may not be suitable in some contexts where the decision maker is also concerned with the magnitude of those errors as well as the "average worst-case" behavior.

B. CONNECTIONS WITH THE LITERATURE

A quantile corresponds to "value-at-risk" (VaR) in financial terminology and relates to "failure probability" in engineering terms. Quantile regression informs the decision maker about these quantities conditional on values of the explanatory random vector X. However, a quantile is not a coherent measure of risk in the sense of Artzner et al. (1999) (see also Delbaen, 2002); it fails to be subadditive. Consequently, a quantile of the sum of two random variables may exceed the sum of the quantiles of each random variable at the same probability level, which runs counter to our understanding of what "risk" should express. Moreover, quantiles cause

computational challenges when incorporated into decision optimization problems as objective function, failure probability constraint, or chance constraint. The use of quantiles and the closely related failure probabilities is therefore problematic in risk-averse decision making; see Artzner et al. (1999), Rockafellar and Uryasev (2000), Rockafellar and Royset (2010), Krokhmal et al. (2011), and Rockafellar and Uryasev (2013) for a detailed discussion.

A superquantile of a random variable, also called "conditional value-at-risk" (CVaR), average value-at-risk, and expected shortfall, is an "average" of certain quantiles as described further below. We prefer the application-neutral name "superquantile" when deriving methods applicable broadly. This is a coherent measure of risk well suited for risk-averse decision making and optimization; see Wang and Uryasev (2007) for its application in financial engineering, Kalinchenko et al. (2011) for military applications, and Rockafellar and Royset (2010) for use in reliability engineering. While this risk measure has reached prominence in risk-averse optimization, there has been much less work on regression techniques that are consistent with it in some sense.

The foundation of least squares and quantile regression is the fact that mean and quantiles minimize the expectation of certain convex random functions. A natural extension to superquantile regression could then possibly involve determining a random function that when minimizing its expectation, we obtain a superquantile. However, such a random function does not exist (as discussed in Gneiting, 2011; Chun et al., 2012), which has led to studies of indirect approaches to superquantile tracking grounded in quantile regression.

For a random variable with a continuous cumulative distribution function, a superquantile equals a conditional expectation of the random variable given realizations no lower than the corresponding quantile. Utilizing this fact, studies have developed kernel-based estimators for the conditional probability density functions, which are then integrated and inverted to obtain estimators of conditional quantiles.

An estimator of the conditional superquantile is then finally constructed by integrating the density estimator over the interval above the quantile (Scaillet, 2005; Cai & Wang, 2008) or forming a sample average (Kato, 2012). These studies also include asymptotic analysis of the resulting estimators under a series of assumptions, including that the data originates from certain time series.

A superquantile of a random variable is defined in terms of an integral of corresponding quantiles with respect to the probability level. Since the integral is approximated by a weighted sum of quantiles across different probability levels, an estimator of a conditional superquantile emerges as the sum of conditional quantiles obtained by quantile regression; see Peracchi and Tanase (2008), and Leorato et al. (2012), which also show asymptotic results under a set of assumptions including the continuous differentiability of the cumulative distribution function of the conditional random variables. Similarly, Chun et al. (2012) utilizes the integral expression for a superquantile, but observes that a weighted sum of quantiles is an optimal solution of a certain minimization problem; see Rockafellar and Uryasev (2013). Analogous to the situation in least squares and quantile regression, an optimization problem yields an estimator of a conditional superquantile. Though, in contrast to the case of least squares and quantile regression, the estimator is "biased" due to the error induced by replacing an integral by a finite sum. Under a linear model assumption, Chun et al. (2012) also constructs a conditional superquantile estimator using an appropriately shifted least squares regression curve based on quantile estimates of residuals. In both cases, asymptotic results are obtained for a homoscedastic linear regression model. Under the same model, Trindade et al. (2007) studies "constrained" regression, where the error random variable $Z_f = Y - f(X)$ is minimized in some sense, for example in terms of least square or absolute deviation, subject to a constraint that limits a superquantile of Z_f. While this approach does not lead to superquantile regression in the sense we derive in this dissertation, it highlights the need for alternative techniques for regression that incorporate superquantiles in some manner.

The need for moving beyond classical regression centered on conditional expectations is now well recognized and has driven even further research towards estimating conditional distribution function, i.e., $P\{Y(x) \leq y\}$ for all $y \in \mathbb{R}$, using nonparametric kernel estimators (see for example Hall & Muller, 2003) and transformation models (see for example Hothorn et al., 2014). We denote by $Y(x)$ the conditional random variable Y given that $X = x \in \mathbb{R}^n$. Of course, conditional distribution functions provide the "full" information about $Y(x)$ including its quantiles and superquantiles, and therefore also provide a means to inform a risk-averse decision maker. In this dissertation, however, we directly focus on superquantiles, which we believe deserve special attention due to their prominence in risk analysis.

A framework for generalized regression is laid out in Rockafellar et al. (2008), and Rockafellar and Uryasev (2013), and regression functions are obtained as optimal solutions of optimization problems of the form $\min_f \mathcal{E}(Z_f)$, where $\mathcal{E}$ is a measure of error and f is restricted to a certain class of functions such as the affine functions. Least squares regression is obtained by $\mathcal{E}(Z_f) = E[Z_f^2]$, quantile regression with the Koenker-Bassett measure of error, but many other possibilities exist. While it is not possible to determine a measure of error that is of the expectation type and yields a superquantile, in Section II.A we show that when allowing for a broader class of functionals, a measure of error that generates a superquantile is indeed available. Such a measure of error is also hinted at in Rockafellar and Royset (2014b), but this dissertation as well as the supporting paper by Rockafellar et al. (2014) gives the first comprehensive treatment. In contrast to previous studies towards superquantile tracking, which utilize indirect approaches and quantile regression, we here offer a natural extension of least squares and quantile regression. We replace the mean-squares and Koenker-Bassett (cf. eq. (II.9)) error measures by a new error measure, and then simply minimize that error of Z_f to obtain a regression function. Under few assumptions, we establish the existence of a regression function, discuss its uniqueness, and examine stability under perturbations of the distribution of (X, Y)

for example caused by sampling. We omit a discussion of simple linear models with independent and identically distributed (iid) noise as we believe that there is little need for quantile and superquantile regression in such contexts as least squares regression with an appropriate shift suffices. In fact, we do not separate models into (additive) deterministic and stochastic terms. In many applications, especially in the area of uncertainty quantification, heteroscedasticity and dependence are prevalent making linear iid and additive models of little value.

C. SCOPE OF DISSERTATION

In this dissertation, we focus on two distinct situations where the importance of a novel regression methodology becomes apparent. We consider a loss random variable Y for which there is incomplete distributional information available, and an explanatory random variable X that is more accessible in some sense.

We denote the first situation and the one we address more often during this dissertation by surrogate estimation. It usually occurs when the explanatory random variable is beyond our direct control, but the dependence between the loss and the explanatory random variable makes us hopeful that, for a carefully selected regression function, such explanatory random variable may serve as a surrogate for the loss random variable. When the distribution of the explanatory random variable is known, at least approximately, and the regression function has been determined, then the distribution of $f(X)$ is usually easily accessible. That distribution may then serve as input to further analysis, simulation, and optimization in place of the unknown distribution of the loss random variable Y. Such surrogate estimation may arise in numerous contexts. "Factor models" in financial investment applications are a result of surrogate estimation (see for example Connor, 1995; Knight et al., 2005), where the random variable we aim to estimate may be the loss associated with a particular asset and the explanatory variable a vector describing a small number of macroeconomic "factors." "Uncertainty quantification" (see for example Lee & Chen,

2009; Eldred et al., 2011) considers the output of a system described by a random variable, for example measuring damage, and estimates its moments and distribution from observed realizations as well as knowledge about the distribution of the input to the system characterized by an explanatory random vector. A main approach here centers on surrogate estimation with the obtained regression function serving as an estimate of the loss random variable.

Another situation arises when the primary concern is with the conditional loss given that the explanatory random variable X takes on specific values. We aim to select these values judiciously in an effort to minimize the conditional loss. We denote this second situation by superquantile tracking. Of course, "minimizing" $Y(x)$ is not well-defined and a standard approach is to minimize a risk measure of $Y(x)$; see for example Krokhmal et al. (2011), and Rockafellar and Uryasev (2013). An attractive choice is to use a superquantile measure of risk, which has nice properties and is also computationally approachable. While in some contexts a superquantile of the conditional loss can be evaluated easily for any specific value of the explanatory random vector, there are numerous situations, especially beyond the financial domain, where only a table of realizations of conditional loss is available for various values of the explanatory random vector. In the latter situation, there is a need for building an approximating model, based on the data, for the relevant superquantile of the conditional loss as a function of the explanatory variables.

D. CONTRIBUTIONS

The main contribution of this dissertation is the development of a novel regression framework that naturally extends least squares and quantile regressions to contexts where one seeks to assess regression errors not by squaring them, as in the case of least squares regression, or by looking at their signs, as in the case of quantile regression, but by weighing larger levels of underestimation increasingly heavily in a manner consistent with superquantiles.

This generalized regression technique is the first attempt to use superquantiles directly in the regression model as opposed to an approximation of conditional quantiles. We develop the fundamental theory for the new regression technique and deal with issues encountered in any generalized regression framework, such as existence and possible uniqueness of the obtained regression functions. We discuss consistency and stability of these regression functions under perturbations due to possible measurement errors and approximating empirical distributions generated by samples of the underlying distribution. And we also examine rate of convergence results under mild assumptions. We present means of assessing the goodness of fit of the obtained quantile and superquantile regression models, by applying the concepts of coefficient of determination, adjusted coefficient of determination, and Cook's distance to quantile and superquantile regression techniques.

We develop two distinct classes of computational methods, one solving the superquantile regression problem by means of analytical and numerical integration techniques, another by relying on the dualization of risk as a step to build a new regression problem that we apply to discrete cases. We discuss complexity results of both classes of computational methods, and compare them to the complexity results for least squares and quantile regressions.

We present a series of numerical examples from the areas of financial investment, military applications, reliability engineering, and uncertainty quantification.

E. DISCLAIMER

The information presented and views expressed in this dissertation do not reflect the official policy or position of the U.S. Navy, the U.S. Department of Defense, the U.S. Government, the Portuguese Navy, and the Portuguese Ministry of National Defense or the Portuguese Government. The data sets we use in our two military applications numerical examples are obtained from unclassified sources, and are employed in this dissertation in order to illustrate some interesting and meaningful

conclusions from our theoretical results.

The first military application example considers the results of an online survey of winged Naval Helicopter Pilots of the U.S. Navy; see Phillips (2011) for details. As stated in Phillips (2011), this study is approved by the NPS Institutional Review Board (IRB) and has an IRB protocol number: NPS.2011.0053-IR-EP7-A. The second military application example considers a data set provided by the Portuguese Navy Submarine Squadron.

F. ORGANIZATION

Chapter II addresses the foundations of the superquantile regression, as an extension of least squares and quantile regressions. The chapter discusses the superquantile regression problem, the issues encountered in such generalized regression frameworks, and provides an approach for assessing the goodness of fit of the obtained quantile and superquantile regression models.

Chapter III develops two classes of computational methods to solve superquantile regression problems. The first denoted by primal method solves superquantile regression problems using analytical and numerical integration schemes. The second which we call the dual method is based on the dualization of risk and utilizes such advantages to build a new superquantile regression problem with promising computational performance, especially for large sample sizes. It also discusses complexity results for the presented algorithms.

Chapter IV provides several numerical results that illustrate not only the primal and dual methods, but also some of the main applications of the superquantile regression, such as superquantile tracking and surrogate estimation.

Chapter V summarizes the theoretical and numerical results, presents our conclusions and suggests future research opportunities.

THIS PAGE INTENTIONALLY LEFT BLANK

II. FOUNDATIONS OF SUPERQUANTILE REGRESSION

In this chapter, we develop a regression technique that extends least squares and quantile regressions, centered on expectations and quantiles, respectively, to one that focuses on superquantiles. This material to a large extent is based on Rockafellar et al. (2014).

Section II.A describes measures of error, risk, deviation, and regret, first in the context of quantile regression and then for the extension to superquantile regression. Section II.B defines superquantile regression as the minimization of a measure of error, provides an alternative approach for solving superquantile regression problems based on the measure of deviation, discusses existence and uniqueness of the regression function, and provides asymptotic results. Section II.C proposes an approach for assessing the goodness of fit of the regression function obtained by quantile and superquantile regressions, using extensions of the definitions of coefficient of determination and Cook's distance.

A. QUANTILES, SUPERQUANTILES, AND ERRORS

While our development centers on superquantiles, it is beneficial to maintain a parallel description of quantiles. As we will see in Subsection II.A.4, quantile regression achieved by minimizing a Koenker-Bassett error of the random variable Z_f, as seen in Subsection II.A.3 in more detail, provides a road map for the construction of superquantile regression, which is simply achieved by minimizing another measure of error. We start, however, with definitions and assumptions, and then provide an overview of the fundamental risk quadrangle, its application to the superquantile as the statistic, and finally we present the corresponding measures of error, deviation, and regret of quantiles and superquantiles.

1. Definitions and Assumptions

We consider a loss random variable Y as a function on a probability space $(\Omega, \mathcal{F}, P)$, and in our context, we assume that Y has a finite second moment, as follows

$$Y \in \mathcal{L}^2 := \mathcal{L}^2(\Omega, \mathcal{F}, P) := \{Y : \Omega \to I\!R \mid Y \text{ is } \mathcal{F}\text{-measurable}, E[Y^2] < \infty\}. \quad \text{(II.1)}$$

Here Ω is a sample space with $\omega \in \Omega$ being a possible outcome; $\mathcal{F}$ is an event space; and P is a probability measure that assigns probabilities to these events, $P : \mathcal{F} \to [0, 1]$.

We now give some useful definitions. We consider the following distinct functionals on $\mathcal{L}^2$, in the sense of Rockafellar and Uryasev (2013), that assign numerical values to random variables, e.g., a loss random variable Y. A *measure of error* $\mathcal{E}(Y)$ quantifies the "nonzeroness" in Y. The $\mathcal{L}^2$-norm of Y is a possible measure of error. A *measure of risk* $\mathcal{R}(Y)$ serves as surrogate for the overall loss in Y. For example, one could think of $\mathcal{R}(Y) = \sup\{Y\}$ (the essential supremum) as such a surrogate, or less conservatively $\mathcal{R}(Y) = E[Y]$. A *measure of deviation* $\mathcal{D}(Y)$ quantifies the "nonconstancy" as uncertainty in Y, and can be seen as a generalization of the standard deviation of Y. A *measure of regret* $\mathcal{V}(Y)$ quantifies the displeasure of obtaining mix realizations of Y, which might be better when $Y \leq 0$ (representing "gains") or worse when $Y > 0$ (representing "losses"). And a *statistic* $\mathcal{S}(Y)$ is associated with Y through $\mathcal{E}$ and $\mathcal{V}$, as described below.

According to Rockafellar et al. (2008), we say that a measure of risk is coherent if the following axioms hold:

(i) $\mathcal{R}(c) = c$ for a constant c.

(ii) $\mathcal{R}(\lambda Y) = \lambda \mathcal{R}(Y)$ when $\lambda > 0$ (positive homogeneity).

(iii) $\mathcal{R}(Y + Y') \leq \mathcal{R}(Y) + \mathcal{R}(Y')$ (subadditivity).

(iv) $\mathcal{R}(Y) \leq \mathcal{R}(Y')$ when $Y \leq Y'$ (monotonicity).

This definition is equivalent to the one described in Artzner et al. (1999), where axiom (i) is replaced by translation invariance. When we refer to a *coherent measure of risk*, we refer to the axioms listed above. The concept of a coherent measure of risk is important in our context because it follows the natural way we think about risk, where monotonicity is a requirement. Moreover, if $\mathcal{R}(Y) > E[Y]$, for a nonconstant random variable Y, then $\mathcal{R}(\cdot)$ is averse.

According to Rockafellar and Uryasev (2013), a *regular measure of risk* satisfies the axiom (i) stated previously, as well as convexity, aversity, and closedness, $\{Y|\mathcal{R}(Y) \leq c\}$ for all constants $c \in I\!R$. Obviously, the expectation is not averse, therefore not regular.

Examples of measures of risk are quantiles and superquantiles of a loss random variable Y at distinct probability levels α, as we define below. For a probability level $\alpha \in (0,1)$, the *α-quantile* of a random variable Y with cumulative distribution function F_Y is defined as

$$q_\alpha(Y) := \min\{y \in I\!R \mid F_Y(y) \geq \alpha\}.$$

Its quantiles are as fundamental to Y as the distribution function, but are problematic to incorporate in risk analysis and optimization due to their lack of coherency as well as increased computational challenges; see Rockafellar and Royset (2014b). Superquantiles have more favorable properties. For $\alpha \in [0,1)$, the *α-superquantile* of a random variable Y is defined as

$$\bar{q}_\alpha(Y) := \frac{1}{1-\alpha}\int_\alpha^1 q_\beta(Y)\, d\beta. \tag{II.2}$$

Since a superquantile is a coherent measure of risk (see Rockafellar & Uryasev, 2000; Rockafellar & Uryasev, 2002) and by virtue of being an "average" of quantiles, it is also more stable than a quantile in some sense, and is well suited for applications. For $\alpha = 1$, we define $\bar{q}_\alpha(Y) := \sup\{Y\}$. Since

$$\bar{q}_0(Y) = \int_0^1 q_\beta(Y) d\beta = \int_0^1 F_Y^{-1}(\beta) d\beta = E[Y], \tag{II.3}$$

we therefore focus on $\alpha \in (0,1)$ throughout the dissertation to avoid distractions by these special cases.

Equivalent to equation (II.2), we have an even more stable and conservative measure of risk, the *α-second-order superquantile*, and it is defined as

$$\bar{\bar{q}}_\alpha(Y) := \frac{1}{1-\alpha}\int_\alpha^1 \bar{q}_\beta(Y)d\beta, \tag{II.4}$$

for a random variable $Y \in \mathcal{L}^2$ and $\alpha \in (0,1)$.

In reliability terminology, quantiles and superquantiles correspond to failure and buffered failure probabilities. The *failure probability* of a loss random variable Y is

$$p(Y) := P\{Y > 0\} = 1 - F_Y(0),$$

which corresponds to

$$p(Y) = 1-\alpha \text{ with } \alpha \text{ such that } q_\alpha(Y) = 0$$

if there is no probability atom at zero. Analogously to the latter expression, the *buffered failure probability* (see Rockafellar & Royset, 2010) of a loss random variable Y is defined as

$$\bar{p}(Y) := 1-\alpha \text{ with } \alpha \text{ such that } \bar{q}_\alpha(Y) = 0. \tag{II.5}$$

Requiring that $\bar{p}(Y) \leq 1-\alpha$ is therefore equivalent to the constraint $\bar{q}_\alpha(Y) \leq 0$. Consequently, in applications with a buffered failure probability constraint on a (conditional) loss random variable $Y(x)$ as well as when the goal is to minimize a superquantile of $Y(x)$ directly, there is a need to estimate $\bar{q}_\alpha(Y(x))$ as a function of $x \in \mathbb{R}^n$. Quantiles and superquantiles are connected through a trade-off formula that leads to quantile regression, as discussed in Subsection II.A.3.

2. Overview of the Fundamental Risk Quadrangle

The "Fundamental Risk Quadrangle" is a concept introduced by Rockafellar and Uryasev (2013), which establishes the connections between distinct measures, described in Subsection II.A.1, of a random variable whose orientation is such that

upper-tail realizations are unfortunate and low realizations are favorable, as described in Chapter I. The interrelationships of such numerical quantities allow distinct comparisons and applications in various analyses, such as risk management.

Diagram 3 in Rockafellar and Uryasev (2013) defines the general relationships between five properties of a random variable Y, measures of error, risk, deviation, and regret, and the corresponding statistic, as we list below. We use these general relationships in the next two subsections.

$$\text{Error measure} = \mathcal{E}(Y) = \mathcal{V}(Y) - E[Y]$$
$$\text{Risk measure} = \mathcal{R}(Y) = \min_{c_0}\{c_0 + \mathcal{V}(Y - c_0)\}$$
$$\text{Deviation measure} = \mathcal{D}(Y) = \mathcal{R}(Y) - E[Y]$$
$$\text{Regret measure} = \mathcal{V}(Y) = \mathcal{E}(Y) + E[Y]$$
$$\text{Statistic} = \mathcal{S}(Y) = \text{argmin}_{c_0}\{c_0 + \mathcal{V}(Y - c_0)\} = \text{argmin}_{c_0}\{\mathcal{E}(Y - c_0)\}$$

We now look at the families of risk quadrangles where the expectation and the quantile are the statistic. The following two risk quadrangles are described in detail in Rockafellar and Uryasev (2013). We list both quadrangles for illustration and to exemplify how one obtains least squares and quantile regressions by minimizing a certain measure of error.

Variance Version of Mean-based Quadrangle:
(Example 1' in Rockafellar & Uryasev, 2013)

$$\text{Error measure} = \mathcal{E}(Y) = \lambda E[Y^2]$$
$$\text{Risk measure} = \mathcal{R}(Y) = E[Y] + \lambda\sigma^2(Y)$$
$$\text{Deviation measure} = \mathcal{D}(Y) = \lambda\sigma^2(Y)$$
$$\text{Regret measure} = \mathcal{V}(Y) = E[Y] + \lambda E[Y^2]$$
$$\text{Statistic} = \mathcal{S}(Y) = E[Y] = \text{mean}$$

Quantile-based Quadrangle: (at any probability level $\alpha \in (0,1)$)
(Example 2 in Rockafellar & Uryasev, 2013)

Error measure $= \mathcal{E}_\alpha(Y) = E\left[\frac{\alpha}{1-\alpha}\max\{0,Y\} + \max\{0,-Y\}\right]$

Risk measure $= \mathcal{R}_\alpha(Y) = \bar{q}_\alpha(Y) = \alpha$-superquantile

Deviation measure $= \mathcal{D}_\alpha(Y) = \bar{q}_\alpha(Y) - E[Y]$

Regret measure $= \mathcal{V}_\alpha(Y) = \frac{1}{1-\alpha}E\left[\max\{0,Y\}\right]$

Statistic $= \mathcal{S}(Y) = q_\alpha(Y) = \alpha$-quantile

With the idea in mind that one minimizes a measure of error to obtain its corresponding statistic in the sense of the "Fundamental Risk Quadrangle," we realize that this approach allows us to naturally extend the existing foundations of least squares and quantile regressions to create new foundations for superquantile regression.

3. Quantile Regret and Error Measures

Both α-quantiles and α-superquantiles, with $\alpha \in (0,1)$, of a loss random variable Y are expressed in terms of an optimization problem involving the measure of regret

$$\mathcal{V}_\alpha(Y) := \frac{1}{1-\alpha}E[\max\{0,Y\}],$$

as seen in Rockafellar and Uryasev (2013). Quantiles and superquantiles then follow as

$$q_\alpha(Y) \in \underset{c_0 \in R}{\operatorname{argmin}}\left\{c_0 + \mathcal{V}_\alpha(Y - c_0)\right\} \tag{II.6}$$

$$\bar{q}_\alpha(Y) = \min_{c_0 \in R}\left\{c_0 + \mathcal{V}_\alpha(Y - c_0)\right\}, \tag{II.7}$$

where in fact $q_\alpha(Y)$ is the lowest optimal solution if multiple solutions exist.

The expression for $q_\alpha(Y)$ is the essential building block for quantile regression, but since we ultimately wish to go beyond the class of constant functions as candidates

for a regression function we need to pass to a measure of error $\mathcal{E}_\alpha$ constructed from $\mathcal{V}_\alpha$ by setting

$$\mathcal{E}_\alpha(Y) := \mathcal{V}_\alpha(Y) - E[Y]$$

for any loss random variable Y (with $E[|Y|] < \infty$). Direct application of the definition of the measure of error and recognition that a constant term in an objective function is immaterial with respect to the optimal solution gives

$$q_\alpha(Y) \in \operatorname*{argmin}_{c_0 \in R} \mathcal{E}_\alpha(Y - c_0), \tag{II.8}$$

with

$$\begin{aligned} \mathcal{E}_\alpha(Y - c_0) &= \frac{1}{1-\alpha} E[\max\{0, Y - c_0\}] - E[Y - c_0] \\ &= E\left[\frac{\alpha}{1-\alpha} \max\{0, Y - c_0\} + \max\{0, -Y + c_0\}\right] \end{aligned} \tag{II.9}$$

being a (scaled) Koenker-Bassett error (Koenker, 2005). Quantile regression centers on computing this argmin with "minimizing the error of $Y - c_0$ over $c_0 \in \mathbb{R}$" replaced by "minimizing the error of $Y - f(X)$ over a class of functions $f : \mathbb{R}^n \to \mathbb{R}$," often taken to be the affine functions. We view $q_\alpha(Y)$ as the "closest" scalar to the random variable Y under a Koenker-Bassett error.

If our goal simply were to estimate $\bar{q}_\alpha(Y)$ of a loss random variable Y for a given probability level $\alpha \in (0, 1)$, the above expressions would have sufficed, possibly passing to an empirical distribution given by a sample if F_Y is unknown. In the present context, however, connections with the underlying explanatory random vector X and the focus on the "approximation" of Y warrants a parallel development to that of quantile regression but now centered on a superquantile. In view of the above review of quantile regression, it is clear that superquantile regression will involve the minimization of some measure of error that returns the superquantile as argmin. Classical least squares regression can be viewed similarly as returning a (conditional) expectation as argmin when minimizing the mean-square measure of error, i.e., $E[Y] = \operatorname{argmin}_{c_0 \in R} E[(Y - c_0)^2]$. The next subsection develops such a measure

of error by first constructing a corresponding measure of regret, for the superquantile as the statistic.

4. Superquantile Regret and Error Measures

We start this subsection by establishing the finiteness of a superquantile under the assumption that the loss random variable Y has a finite second moment.

We know from Rockafellar and Uryasev (2013) that $\bar{q}_\alpha$ is a convex, positively homogenous, monotonic, and averse functional on $\mathcal{L}^2$ for $\alpha \in (0,1)$. From Theorem 3 in Rockafellar and Royset (2014b), see also Rockafellar et al. (2014), we know that $\bar{q}_\alpha$ is bounded as stated next. We adopt the notation $\sigma^2(Y) = E[(Y - E[Y])^2]$.

Proposition II.1. *For $Y \in \mathcal{L}^2$ and $\alpha \in (0,1)$ one has that*

$$\bar{q}_\alpha(Y) \leq E[Y] + \frac{1}{\sqrt{1-\alpha}}\sigma(Y). \tag{II.10}$$

Proof.

Suppose that the quantile $q_\alpha(Y)$, viewed as a function of the probability level, is continuous at α. Let I_α be the indicator function of the interval $[q_\alpha(Y), \infty)$. We then have by the Schwarz inequality that

$$\begin{aligned}(1-\alpha)\bar{q}_\alpha(Y - E[Y]) &= E[(Y - E[Y])I_\alpha] \\ &\leq \sqrt{E[(Y - E[Y])^2]}\sqrt{E[I_\alpha^2]} \\ &\leq \sigma(Y)\sqrt{1-\alpha}.\end{aligned}$$

Then, since $\bar{q}_\alpha(Y - E[Y]) = \bar{q}_\alpha(Y) - E[Y]$, the result follows from dividing by $1-\alpha$. Thus, (II.10) is valid under the continuity assumption about the quantile, which is true for all but at most countable many α. By continuity on both sides of (II.10) with respect to α, it must then hold for all $\alpha \in (0,1)$.

□

The measure of regret at probability level $\alpha \in (0,1)$ that serves in the context of superquantile regression is defined for any loss random variable Y as

$$\bar{\mathcal{V}}_\alpha(Y) := \frac{1}{1-\alpha}\ \bar{\mathcal{V}}_0(Y), \tag{II.11}$$

where

$$\bar{\mathcal{V}}_0(Y) := \int_0^1 \max\{0, \bar{q}_\beta(Y)\} d\beta. \tag{II.12}$$

These expressions appear in Rockafellar and Royset (2014b), where their discovery, which is related to the Hardy-Littlewood transform, is described. Here, we provide the alternative, direct proof of Rockafellar et al. (2014), on how these expressions lead to the superquantile as optimal solution of (II.7). We start, however, with two preliminary results and the definition of a corresponding measure of error.

Lemma II.1. *For $Y \in \mathcal{L}^2$,*

$$\bar{\mathcal{V}}_0(Y) \leq \sigma(Y) + \max\{0, E[Y] + \sigma(Y)\}. \tag{II.13}$$

Proof.

From (II.10) and (II.12) we have

$$\bar{\mathcal{V}}_0(Y) \leq \int_0^1 \max\{0, \theta_Y(\beta)\} d\beta \quad \text{for} \;\; \theta_Y(\beta) = E[Y] + \frac{1}{\sqrt{1-\beta}}\sigma(Y). \tag{II.14}$$

We consider three cases. In Case 1, we suppose that $\theta_Y(\beta) \geq 0$ for all $\beta \in [0, 1]$. Then the right hand side of (II.14) is given by

$$\int_0^1 \theta_Y(\beta) d\beta = E[Y] + \sigma(Y) \int_0^1 (1-\beta)^{-1/2} d\beta \quad \text{with} \int_0^1 (1-\beta)^{-1/2} d\beta = 2. \tag{II.15}$$

Therefore, $\bar{\mathcal{V}}_0(Y) \leq E[Y] + 2\sigma(Y)$ in Case 1. In Case 2a, we suppose that $\theta_Y(\beta) \leq 0$ for all $\beta \in (0, 1)$. Then obviously $\bar{\mathcal{V}}_0(Y) \leq 0$. Finally, in Case 2b, let $\theta_Y(\beta) < 0$ for some $\beta \in (0, 1)$, but not all. Then necessarily $\sigma(Y) > 0$ and $E[Y] \leq -\sigma(Y)$, and $\theta_Y(\beta)$ strictly increases with respect to β. Let $\bar{\alpha}$ be the unique $\beta \in (0, 1)$ with

$\theta_Y(\bar{\alpha}) = 0$, namely when $\sqrt{1-\bar{\alpha}} = \sigma(Y)/(-E[Y])$. Then we have that

$$\begin{aligned}
\int_0^1 \max\{0, \theta_Y(\beta)\} d\beta &= \int_{\bar{\alpha}}^1 \theta_Y(\beta) d\beta \\
&= (1-\bar{\alpha})E[Y] + \sigma(Y) \int_{\bar{\alpha}}^1 (1-\beta)^{-1/2} d\beta \\
&= (1-\bar{\alpha})E[Y] + 2\sigma(Y)\sqrt{1-\bar{\alpha}} \\
&= \frac{\sigma(Y)^2}{E[Y]^2} E[Y] + 2\sigma(Y) \frac{\sigma(Y)}{-E[Y]} \\
&= \frac{\sigma(Y)^2}{-E[Y]} \\
&\leq \sigma(Y).
\end{aligned}$$

Thus, in Case 2b we get $\bar{\mathcal{V}}_0(Y) \leq \sigma(Y)$. The conclusion then follows by putting together the cases.

□

We observe that for $\alpha \in (0,1)$, $\bar{\mathcal{V}}_\alpha$ is also a convex, positively homogeneous, monotonic, and averse functional on $\mathcal{L}^2$, which follows from the properties of the superquantile (Rockafellar & Uryasev, 2013), and by the above result it is also finite, and consequently continuous. A corresponding measure of error is defined for $Y \in \mathcal{L}^2$ by

$$\bar{\mathcal{E}}_\alpha(Y) := \bar{\mathcal{V}}_\alpha(Y) - E[Y] \tag{II.16}$$

and referred to as a superquantile error. Obviously, $\bar{\mathcal{E}}_\alpha$ is also convex and positively homogeneous. It also satisfies the following properties.

Proposition II.2. *For any $\alpha \in (0,1)$ and $Y \in \mathcal{L}^2$, a superquantile error satisfies*

(a) $\bar{\mathcal{E}}_\alpha(Y) = 0$ *when* $Y \equiv 0$,

(b) $\bar{\mathcal{E}}_\alpha(Y) > 0$ *when* $Y \not\equiv 0$, *and*

(c) $\bar{\mathcal{E}}_\alpha(Y) \geq \min\{1, \alpha/(1-\alpha)\}|E[Y]|$.

Proof.

Since $\bar{q}_\beta(0) = 0$ for all $\beta \in [0,1]$, (a) follows trivially.

Since $\bar{\mathcal{V}}_\alpha$ is averse, we have that for $Y \in \mathcal{L}^2$ $\bar{\mathcal{E}}_\alpha(Y) = \bar{\mathcal{V}}_\alpha(Y) - E[Y] > E[Y] - E[Y] = 0$ when Y is not a constant. To complete part (b), we therefore only need to consider nonzero constants. If Y is a positive constant K, then

$$\begin{aligned}\frac{1}{1-\alpha}\int_0^1 \max\{0, \bar{q}_\beta(Y)\}d\beta - E[Y] &> \int_0^1 \max\{0, \bar{q}_\beta(Y)\}d\beta - E[Y] \\ &> K - E[Y] \\ &> 0.\end{aligned}$$

If Y is a negative constant K, then

$$\begin{aligned}\frac{1}{1-\alpha}\int_0^1 \max\{0, \bar{q}_\beta(Y)\}d\beta - E[Y] &= \frac{1}{1-\alpha}\int_0^1 \max\{0, K\}d\beta - E[Y] \\ &= 0 - E[Y] \\ &> 0,\end{aligned}$$

which completes part (b).

Since $\bar{q}_\beta(Y) \geq E[Y]$ for all $\beta \in [0,1]$, we have whenever $E[Y] \geq 0$ the bound

$$\begin{aligned}\frac{1}{1-\alpha}\int_0^1 \max\{0, \bar{q}_\beta(Y)\}d\beta - E[Y] &\geq \frac{1}{1-\alpha}\int_0^1 \max\{0, E[Y]\}d\beta - E[Y] \\ &\geq \frac{\alpha}{1-\alpha}E[Y].\end{aligned}$$

And when $E[Y] < 0$,

$$\begin{aligned}\frac{1}{1-\alpha}\int_0^1 \max\{0, \bar{q}_\beta(Y)\}d\beta - E[Y] &\geq \frac{1}{1-\alpha}\int_0^1 \max\{0, E[Y]\}d\beta - E[Y] \\ &\geq -E[Y].\end{aligned}$$

Part (c) then follows by combining the two results.

□

By Proposition II.2 and the above discussion, $\bar{\mathcal{E}}_\alpha$ is a regular measure of error. We now show that a superquantile is a unique optimal solution of optimization problems involving $\bar{\mathcal{V}}_\alpha$ and $\bar{\mathcal{E}}_\alpha$. As mentioned, the connection between a superquantile and $\bar{\mathcal{V}}_\alpha$ is also reached in Theorem 7 of Rockafellar and Royset (2014b) through different means. Here we derive the direct proof and the connection with a superquantile error (see Rockafellar et al., 2014).

Theorem II.1. *(Superquantile as optimal solution) For $Y \in \mathcal{L}^2$ and $\alpha \in (0,1)$,*

$$\begin{aligned} \bar{q}_\alpha(Y) &= \operatorname*{argmin}_{c_0 \in R} \{c_0 + \bar{\mathcal{V}}_\alpha(Y - c_0)\} \\ &= \operatorname*{argmin}_{c_0 \in R} \bar{\mathcal{E}}_\alpha(Y - c_0). \end{aligned} \tag{II.17}$$

Proof.

Let $\varphi(c) = c + \bar{\mathcal{V}}_\alpha(Y - c)$ and $\psi_\beta(c) = \max\{0, \bar{q}_\beta(Y) - c\}$. These are both convex functions of c, and ψ_β is nonincreasing. We can use the criterion that

$$\bar{c} \in \operatorname*{argmin}_c \varphi(c) \Longleftrightarrow \varphi'_+(\bar{c}) \geq 0,\ \varphi'_-(\bar{c}) \leq 0,$$

where, because of the monotonicity of ψ_β,

$$\varphi'_+(c) = 1 + \frac{1}{1-\alpha}\int_0^1 (\psi_\beta)'_-(c)d\beta, \qquad \varphi'_-(c) = 1 + \frac{1}{1-\alpha}\int_0^1 (\psi_\beta)'_+(c)d\beta,$$

$$(\psi_\beta)'_+(c) = \begin{cases} -1 & \text{if } \bar{q}_\beta(Y) > c, \\ 0 & \text{if } \bar{q}_\beta(Y) \leq c, \end{cases} \qquad (\psi_\beta)'_-(c) = \begin{cases} -1 & \text{if } \bar{q}_\beta(Y) \geq c, \\ 0 & \text{if } \bar{q}_\beta(Y) < c. \end{cases}$$

Therefore

$$\begin{aligned} \int_0^1 (\psi_\beta)'_+(c)d\beta &= \int_0^1 (\psi_\beta)'_-(c)d\beta \\ &= -(1-\gamma) \qquad \text{for } c = \bar{q}_\gamma(Y), \end{aligned}$$

in which case $(\psi_\beta)'(c) = (\psi_\beta)'_+(c) = (\psi_\beta)'_-(c) = 1-(1-\gamma)/(1-\alpha)$. Thus, $(\psi_\beta)'(c) = 0$ corresponds to $c = \bar{q}_\gamma(Y)$ for $\gamma = \alpha$. Consequently, the first equality of the theorem holds. The second follows directly from (II.16) and the fact that a constant in an objective function is immaterial with regard to the argmin.

□

The foundations for quantile regression are given by equations (II.6) and (II.8). Analogously, the expressions in (II.17) provide the path to superquantile regression as developed in Section II.B. In fact, Theorem II.1 shows that $\bar{q}_\alpha(Y)$ is the uniquely "closest" scalar to Y in the sense of the superquantile error. The optimal value in (II.17) defines a measure of risk (see Rockafellar & Royset, 2014b)

$$\begin{aligned} \bar{\mathcal{R}}_\alpha(Y) &:= \min_{c_0 \in R} \{c_0 + \bar{\mathcal{V}}_\alpha(Y - c_0)\} \\ &= \bar{q}_\alpha(Y) + \bar{\mathcal{V}}_\alpha(Y - \bar{q}_\alpha(Y)) \end{aligned}$$

for $Y \in \mathcal{L}^2$ analogously to $\bar{q}_\alpha(Y)$ in (II.7). A corresponding measure of deviation is given by

$$\begin{aligned} \bar{\mathcal{D}}_\alpha(Y) &:= \min_{c_0 \in \boldsymbol{R}} \bar{\mathcal{E}}_\alpha(Y - c_0) \\ &= \bar{\mathcal{R}}_\alpha(Y) - E[Y]. \end{aligned} \tag{II.18}$$

We note that parallel to (II.2) (see Rockafellar & Royset, 2014b),

$$\bar{\mathcal{R}}_\alpha(Y) = \frac{1}{1-\alpha} \int_\alpha^1 \bar{q}_\beta(Y) d\beta$$

and, consequently,

$$\bar{\mathcal{D}}_\alpha(Y) = \frac{1}{1-\alpha} \int_\alpha^1 \bar{q}_\beta(Y) d\beta - E[Y].$$

The measures of regret, error, risk, and deviation, $\bar{\mathcal{V}}_\alpha$, $\bar{\mathcal{E}}_\alpha$, $\bar{\mathcal{R}}_\alpha$, and $\bar{\mathcal{D}}_\alpha$, respectively, for a probability level $\alpha \in (0,1)$, form a family of risk quadrangles, in the sense of Rockafellar and Uryasev (2013), that corresponds to the superquantile as the statistic, as shown below.

Superquantile-based Quadrangle: (at any probability level $\alpha \in (0,1)$)

Error measure $= \bar{\mathcal{E}}_\alpha(Y) = \frac{1}{1-\alpha} \int_0^1 \max\{0, \bar{q}_\beta(Y)\}\, d\beta - E[Y]$

Risk measure $= \bar{\mathcal{R}}_\alpha(Y) = \bar{\bar{q}}_\alpha(Y) = \alpha$-second-order superquantile

Deviation measure $= \bar{\mathcal{D}}_\alpha(Y) = \bar{\bar{q}}_\alpha(Y) - E[Y]$

Regret measure $= \bar{\mathcal{V}}_\alpha(Y) = \frac{1}{1-\alpha} \int_0^1 \max\{0, \bar{q}_\beta(Y)\}\, d\beta$

Statistic $= \mathcal{S}(Y) = \bar{q}_\alpha(Y) = \alpha$-superquantile

We note here that the measure of deviation $\bar{\mathcal{D}}_\alpha$ plays a central role in the remainder of the dissertation as it facilitates simplifications, goodness of fit tests, and computational methods.

B. SUPERQUANTILE REGRESSION

Theorem II.1 and the development leading to quantile regression direct us to a new regression methodology that is centered on a superquantile error. The next subsection poses the regression problem, provides its properties, and discusses stability under perturbations. The section ends with a discussion of superquantile tracking.

1. Superquantile Regression Problem

While Theorem II.1 shows that the "best" scalar approximation of a random variable Y in the sense of a superquantile error is the corresponding superquantile, we now go beyond the class of constant functions to utilize the connection with an underlying explanatory random vector X. We focus on regression functions of the form

$$f(x) = c_0 + \langle c, h(x) \rangle, \qquad c_0 \in I\!R, c \in I\!R^m,$$

for a given "basis" function $h : I\!R^n \to I\!R^m$. This class satisfies most practical needs including that of linear regression where $m = n$ and $h(x) = x$. Extensions beyond this class are also possible but not dealt with in this dissertation.

We now define the Superquantile Regression Problem SqR, for any $h : I\!R^n \to I\!R^m$ and $\alpha \in (0, 1)$, where

$$Z(c_0, c) := Y - (c_0 + \langle c, h(X) \rangle)$$

is the error random variable, whose distribution depends on c_0, c, h, and the joint distribution of (X, Y). We denote by $\bar{\mathcal{C}} \subset I\!R^{m+1}$ the set of optimal solutions of SqR and refer to $(\bar{c}_0, \bar{c}) \in \bar{\mathcal{C}}$ as a regression vector.

Superquantile Regression Problem:

$$SqR: \quad \min_{c_0 \in R, c \in R^m} \bar{\mathcal{E}}_\alpha \left(Z(c_0, c) \right) = \frac{1}{1-\alpha} \int_0^1 \max \left\{ 0, \bar{q}_\beta(Z(c_0, c)) \right\} d\beta - E[Z(c_0, c)].$$

The objective function $\bar{\mathcal{E}}_\alpha(Z(\cdot,\cdot))$ is well-defined and finite when the distribution of (X,Y) and h is such that $Z(c_0,c) \in \mathcal{L}^2$ for all $c_0 \in I\!R$, and $c \in I\!R^m$. A sufficient condition that ensures this property is that $Y,\ h_1(X),...,\ h_m(X) \in \mathcal{L}^2$. We adopt the notation

$$H = h(X),\ H_i = h_i(X),\ i = 1,2,...,m.$$

Lemma II.2. *If $Y,\ H_1,...,\ H_m \in \mathcal{L}^2$, then $Z(c_0,c) \in \mathcal{L}^2$ for all $c_0 \in I\!R$, and $c \in I\!R^m$.*

In surrogate estimation, $\bar{c}_0 + \langle \bar{c}, h(X)\rangle$, with $(\bar{c}_0,\bar{c}) \in \bar{\mathcal{C}}$, provides the best approximation of Y in the sense of a superquantile error. For example, after having computed $(\bar{c}_0,\bar{c})$, the analysis could proceed with examining the moments, quantiles, and superquantiles of $\bar{c}_0 + \langle \bar{c}, h(X)\rangle$ as surrogates for the corresponding quantities of Y. If X is Gaussian and h is affine, then $\bar{c}_0 + \langle \bar{c}, h(X)\rangle$ is a Gaussian approximation of Y easily examined and utilized in further studies. It may also be of interest to examine $\bar{c}_0 + \langle \bar{c}, h(X)\rangle$ under hypothetical distributions of X.

As a direct consequence of the Regression Theorem in Rockafellar and Uryasev (2013) (see also Theorem 3.1 in Rockafellar et al., 2008), we obtain that a regression vector can equivalently be determined from a measure of deviation $\bar{\mathcal{D}}_\alpha$.

Proposition II.3. *Suppose that $Y, H_1, ..., H_m \in \mathcal{L}^2$. Then, the set of regression vectors $\bar{\mathcal{C}}$ of SqR is equivalently obtained as*

$$\bar{\mathcal{C}} = \left\{ (\bar{c}_0,\bar{c}) \in I\!R^{m+1} \mid \bar{c} \in \operatorname*{argmin}_{c \in R^m} \bar{\mathcal{D}}_\alpha(Z_0(c)),\quad \bar{c}_0 = \bar{q}_\alpha(Z_0(\bar{c})) \right\},$$

where $Z_0(c) := Y - \langle c, h(X)\rangle$.

Proposition II.3 implies computational advantages as the $(m+1)$-dimensional optimization problem SqR is replaced by a problem in m dimensions with a simpler objective function, which we fully utilize in Chapters III and IV. Moreover, the result also proves to be beneficial in analysis of regression vectors.

We now define the Deviation-based Superquantile Regression Problem $DSqR$, for any $h : I\!R^n \to I\!R^m$ and $\alpha \in (0, 1)$:

Deviation-based Superquantile Regression Problem:

$$DSqR: \qquad \min_{c \in R^m} \ \bar{\mathcal{D}}_\alpha\left(Z_0(c)\right) = \frac{1}{1-\alpha}\int_\alpha^1 \bar{q}_\beta(Z_0(c))d\beta - E[Z_0(c)],$$

with $\bar{c}_0$ being obtained by setting $\bar{c}_0 = \bar{q}_\alpha(Z_0(\bar{c}))$.

The existence of a regression vector is ensured by the next result, which also provides conditions for uniqueness.

Theorem II.2. *(Existence and uniqueness of regression vector) If $Y, H_1, ..., H_m \in \mathcal{L}^2$, then SqR is a convex problem with a set of optimal solutions $\bar{\mathcal{C}}$ that is nonempty, closed, and convex.*

(a) $\bar{\mathcal{C}}$ is bounded if and only if the random vector X and the basis function h satisfy the condition that $\langle c, h(X)\rangle$ is not constant unless $c = 0$.

(b) If in addition, for every $(c_0, c), (c_0', c') \in I\!R^{m+1}$, with $c \neq c'$, there exists a $\beta_0 \in [0, 1)$ such that

$$0 \leq \bar{q}_\beta(Z(c_0, c) + Z(c_0', c')) < \bar{q}_\beta(Z(c_0, c)) + \bar{q}_\beta(Z(c_0', c')) \tag{II.19}$$

for all $\beta \in [\beta_0, 1)$, then $\bar{\mathcal{C}}$ is a singleton.

Proof.

Since $Y \in \mathcal{L}^2$ implies that $\bar{\mathcal{E}}_\alpha(Y) < \infty$, by Lemma II.1, we deduce the two first conclusions from Theorem 3.1 in Rockafellar et al. (2008). Hence, we only need to show that $\bar{\mathcal{C}}$ is a singleton.

Suppose for the sake of a contradiction that $(c_0, c), (c_0', c') \in \bar{\mathcal{C}}$ and $(c_0, c) \neq (c_0', c')$, with corresponding optimal value $\xi \geq 0$, i.e., $\xi = \bar{\mathcal{E}}_\alpha(Z(c_0, c)) = \bar{\mathcal{E}}_\alpha(Z(c_0', c'))$. We consider two cases.

First, suppose that $\xi = 0$. By Proposition II.2, $Z(c_0, c) = Z(c_0', c') = 0$ and consequently

$$c_0 + \langle c, H\rangle = c_0' + \langle c', H\rangle,$$

which implies that $\langle c - c', H\rangle = c'_0 - c_0$. Under the assumption that $\langle c, h(X)\rangle$ is only constant when $c = 0$, we must have that $c - c' = 0$. Then, also $c'_0 - c_0 = 0$ follows, which contradicts the hypothesis that $(c_0, c) \neq (c'_0, c')$.

Second, suppose that $\xi > 0$. If $c = c'$, then a direct consequence of Proposition II.3 and the fact that every random variable has a unique superquantile at each probability level, is that also $c_0 = c'_0$, which again contradicts our hypothesis. Consequently, we focus on the case with $c \neq c'$, for which there exists a β_0 such that (II.19) holds for all $\beta \in [\beta_0, 1)$. Trivially, then

$$\max\{0, \bar{q}_\beta(Z(c_0, c) + Z(c'_0, c'))\} < \max\{0, \bar{q}_\beta(Z(c_0, c))\} + \max\{0, \bar{q}_\beta(Z(c'_0, c'))\}$$

for $\beta \in [\beta_0, 1)$. If $\beta \in (0, 1)$ is such that $\bar{q}_\beta(Z(c_0, c) + Z(c'_0, c')) < 0$, then

$$\max\{0, \bar{q}_\beta(Z(c_0, c) + Z(c'_0, c'))\} \leq \max\{0, \bar{q}_\beta(Z(c_0, c))\} + \max\{0, \bar{q}_\beta(Z(c'_0, c'))\}$$

as the left-hand side vanishes and the right-hand side is nonnegative. Hence,

$$\int_0^1 \max\{0, \bar{q}_\beta(Z(c_0, c) + Z(c'_0, c'))\} d\beta \\ < \int_0^1 \max\{0, \bar{q}_\beta(Z(c_0, c))\} d\beta + \int_0^1 \max\{0, \bar{q}_\beta(Z(c'_0, c'))\} d\beta$$

and also

$$\bar{\mathcal{E}}_\alpha(Z(c_0, c) + Z(c'_0, c')) < \bar{\mathcal{E}}_\alpha(Z(c_0, c)) + \bar{\mathcal{E}}_\alpha(Z(c'_0, c')). \tag{II.20}$$

Let

$$(c''_0, c'') = (1/2)(c_0, c) + (1/2)(c'_0, c')$$

and therefore

$$2Z(c''_0, c'') = Z(c_0, c) + Z(c'_0, c').$$

By the optimality of ξ, the positive homogeneity of $\bar{\mathcal{E}}_\alpha$, and (II.20), we find that

$$2\xi \leq 2\bar{\mathcal{E}}_\alpha(Z(c''_0, c'')),$$

and that

$$\begin{aligned} 2\bar{\mathcal{E}}_\alpha(Z(c_0'', c'')) &= \bar{\mathcal{E}}_\alpha(2Z(c_0'', c'')) \\ &< \bar{\mathcal{E}}_\alpha(Z(c_0, c)) + \bar{\mathcal{E}}_\alpha(Z(c_0', c')). \end{aligned}$$

Since

$$\bar{\mathcal{E}}_\alpha(Z(c_0, c)) + \bar{\mathcal{E}}_\alpha(Z(c_0', c')) = 2\xi,$$

we finally get that

$$2\xi \leq 2\xi,$$

which cannot hold. In view of this contradiction, the conclusion follows.

□

While Theorem II.2 gives a sufficient condition for uniqueness of the regression vector, in general uniqueness cannot be expected. For example, suppose that the random vector (X, Y), with X scalar valued, has the possible and equally likely realizations $(1, 1)$, $(2, 2)$, and $(3, 1)$. Then, $\bar{q}_\beta(Z_0(c)) = \max\{1 - c, 2 - 2c, 1 - 3c\}$ for $\beta > 2/3$ and $E[Z_0(c)] = 4/3 - 2c$. It is straightforward to show that for $\alpha > 2/3$, any $c \in [-1, 1]$ minimizes $\bar{\mathcal{D}}_\alpha(Z_0(\cdot))$. Consequently, in view of Proposition II.3, any $c \in [-1, 1]$, with a corresponding $c_0 = \max\{1-c, 2-2c, 1-3c\}$, minimizes $\bar{\mathcal{E}}_\alpha(Z(\cdot, \cdot))$ for $\alpha > 2/3$, as shown in Figure 5. The minimum error is $2/3$.

A unique regression vector is indeed achieved in the normal case as stated next.

Proposition II.4. *Suppose that (H, Y) is normally distributed with positive definite variance-covariance matrix. Then, $\bar{\mathcal{C}}$ is a singleton.*

Proof.

Let Σ be the variance-covariance matrix of (H, Y), with Cholesky decomposition $\Sigma = LL^\top$. For any $\beta \in (0, 1)$ and $c \in \mathbb{R}^m$, $Z_0(c)$ is also normal with mean $E[Z_0(c)] = \langle \tilde{c}, E[(H, Y)] \rangle$ and variance $\sigma^2(Z_0(c)) = \langle \tilde{c}, \Sigma \tilde{c} \rangle$, where $\tilde{c} = (-c, 1)$. Thus,

$$\bar{q}_\beta(Z_0(c)) = E[Z_0(c)] + k_\beta \sigma(Z_0(c)) = E[Z_0(c)] + k_\beta \|L^\top \tilde{c}\|,$$

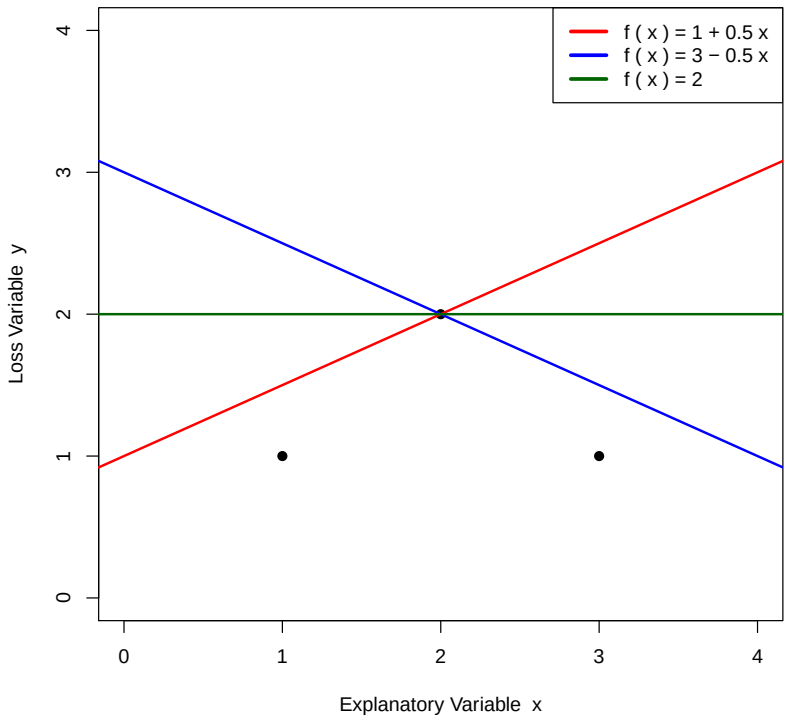


Figure 5. Example of multiple optimal solutions for problem SqR.

where $k_\beta = \phi(\Phi^{-1}(\beta))/(1-\beta)$, with ϕ and Φ being the standard normal probability density and cumulative distribution functions, respectively.

For $c, c' \in I\!R^m$, with $c \neq c'$, there is no constant $k > 0$ such that $(-c, 1) = k(-c', 1)$. Let $\tilde{c} = (-c, 1)$ and $\tilde{c}' = (-c', 1)$. Since Σ is positive definite, the upper-triangular matrix $L^\top$ is unique and full rank. Consequently, the null space of $L^\top$ contains only the zero vector and $L^\top(\tilde{c} - k\tilde{c}') \neq 0$ for all scalars $k > 0$. Since the triangle inequality for two vectors holds strictly whenever the two vectors cannot be expressed as a positive multiple of each other, we therefore find that

$$\|L^\top \tilde{c} + L^\top \tilde{c}'\| < \|L^\top \tilde{c}\| + \|L^\top \tilde{c}'\|.$$

Now for the sake of a contradiction suppose that $c, c' \in I\!R^m$ both minimize $\bar{\mathcal{D}}_\alpha(Z_0(\cdot))$ and attain the minimum value $\xi \in I\!R$, but $c \neq c'$. Let

$$c'' = (1/2)c + (1/2)c', \quad \tilde{c}'' = (-c'', 1), \quad \text{and } \gamma_\alpha = \int_\alpha^1 k_\beta d\beta/(1-\alpha) > 0.$$

Then,

$$\begin{aligned}\bar{\mathcal{D}}_\alpha(Z_0(c'')) &= \frac{1}{1-\alpha}\int_\alpha^1 \bar{q}_\beta(Z_0(c''))d\beta - E[Z_0(c'')] \\ &= E[Z_0(c'')] + \gamma_\alpha\|L^\top\tilde{c}''\| - E[Z_0(c'')] \\ &= \frac{\gamma_\alpha}{2}\|L^\top\tilde{c} + L^\top\tilde{c}'\| \\ &< \frac{\gamma_\alpha}{2}(\|L^\top\tilde{c}\| + \|L^\top\tilde{c}'\|),\end{aligned}$$

and since

$$\begin{aligned}\frac{\gamma_\alpha}{2}(\|L^\top\tilde{c}\| + \|L^\top\tilde{c}'\|) &= \frac{1}{2}\Big(E[Z_0(c)] + \gamma_\alpha\|L^\top\tilde{c}\| - E[Z_0(c)]\Big) + \\ &\qquad + \frac{1}{2}\Big(E[Z_0(c')] + \gamma_\alpha\|L^\top\tilde{c}'\| - E[Z_0(c')]\Big) \\ &= \frac{1}{2}\Big(\bar{\mathcal{D}}_\alpha(Z_0(c))\Big) + \frac{1}{2}\Big(\bar{\mathcal{D}}_\alpha(Z_0(c'))\Big) \\ &= \frac{1}{2}(\xi + \xi) \\ &= \xi,\end{aligned}$$

we have that $\bar{\mathcal{D}}_\alpha(Z_0(c'')) < \xi$. However, this contradicts the optimality of c, c' and we reach the conclusion.

□

We next turn to consistency and stability of the regression vector. Of course, the joint distribution of (X, Y) is rarely available in practice and one may need to pass to an approximating empirical distribution generated by a sample. Moreover, perturbations of the "true" distribution of (X, Y) may occur due to measurement errors in the data and other factors. We consider these possibilities and let (X^ν, Y^ν) be a random vector whose joint distribution approximates that of (X, Y) in some sense. For example, (X^ν, Y^ν) may be governed by the empirical distribution generated by an independent and identically distributed sample of size ν from (X, Y). Presumably, as $\nu \to \infty$, the approximation of (X, Y) by (X^ν, Y^ν) improves as stated formally below. Regardless of the nature of (X^ν, Y^ν), we define the approximate error random variable as

$$Z^\nu(c_0, c) := Y^\nu - c_0 - \langle c, h(X^\nu)\rangle,$$

and the corresponding Approximate Superquantile Regression Problem SqR^ν as follows:

Approximate Superquantile Regression Problem:

$$SqR^\nu: \quad \min_{c_0 \in R, c \in R^m} \ \bar{\mathcal{E}}_\alpha\left(Z^\nu(c_0,c)\right) = \frac{1}{1-\alpha}\int_0^1 \max\left\{0, \bar{q}_\beta(Z^\nu(c_0,c))\right\} d\beta - \\ -E[Z^\nu(c_0,c)].$$

The next result shows that as (X^ν, Y^ν) approximates (X,Y), a regression vector obtained from SqR^ν approximates one from SqR, which provides the justification for basing a regression analysis on SqR^ν. Below, we let $\to^d$ denote convergence in distribution and

$$H^\nu = h(X^\nu),\ H_i^\nu = h_i(X^\nu),\ i = 1,2,...,m.$$

Theorem II.3. *(Stability of regression vector) Suppose that (X^ν, Y^ν), $\nu = 1,2,...$, and (X,Y) are $n+1$-dimensional random vectors such that $(X^\nu, Y^\nu) \to^d (X,Y)$ and that the basis function h is continuous except possibly on a subset $S \subset I\!R^n$ with $P\{X \in S\} = 0$. Moreover, let $H_i, Y \in \mathcal{L}^2$, $\sup_\nu E[(H_i^\nu)^2] < \infty$, $i = 1,2,...,m$, and $\sup_\nu E[(Y^\nu)^2] < \infty$.*

If $\{(\bar{c}_0^\nu, \bar{c}^\nu)\}_{\nu=1}^\infty$ is a sequence of optimal solutions of SqR^ν, with $\alpha \in (0,1)$, then every accumulation point of that sequence is a regression vector of SqR.

Proof.

Let $(c_0, c) \in I\!R^{m+1}$ be arbitrary. By the continuous mapping theorem (see for example Theorem 29.2 in Billingsley, 1995),

$$Z^\nu(c_0,c) = Y^\nu - c_0 - \langle c, h(X^\nu)\rangle \quad \to^d \quad Z(c_0,c) = Y - c_0 - \langle c, h(X)\rangle.$$

By the assumed moment conditions in (II.1), there exists a constant $M < \infty$ that bounds from above the terms

$$\max_i E[|H_i|], \quad \max_i E[(H_i)^2], \quad \sup_{\nu,i} E[|H_i^\nu|], \quad \sup_{\nu,i} E[(H_i^\nu)^2], \\ \text{and} \quad E[|Y|], \quad E[Y^2], \quad \sup_\nu E[|Y^\nu|], \quad \sup_\nu E[(Y^\nu)^2].$$

In view of Lemma II.2 and its proof, we deduce that

$$E[(Y^\nu - c_0 - \langle c, H^\nu\rangle)^2] \leq M + 2(\|c\|m^{1/2}M + (M + |c_0|)\|c\|mM) + \|c\|^2 mM \quad \text{(II.21)}$$

for all ν. Hence, $Z^\nu(c_0, c)$ is uniformly integrable (for fixed c_0, c) and

$$E[Z^\nu(c_0, c)] \quad \rightarrow \quad E[Z(c_0, c)] < \infty; \quad \text{(II.22)}$$

see Billingsley (1995), Theorem 25.12 and its corollary.

By Theorem 4 in Rockafellar and Royset (2014b), a sequence of random variables converges in distribution to a random variable if and only if the corresponding α-superquantiles, viewed as functions of the probability level α, converge uniformly on every closed subset of $(0, 1)$. Consequently, $\bar{q}_\beta(Z^\nu(c_0, c)) \rightarrow \bar{q}_\beta(Z(c_0, c))$ uniformly in β on closed subsets of $(0, 1)$. Moreover, since the 0-superquantile coincides with the expectation, (II.22) implies that $\bar{q}_0(Z^\nu(c_0, c)) \rightarrow \bar{q}_0(Z(c_0, c))$ also holds. These facts and the observation that the superquantile of any random variable is continuous and nondecreasing as a function of the probability level, ensure that for any $\epsilon > 0$ and $\delta \in (0, 1)$, there exists an integer $\nu(\epsilon, \delta)$ such that for all $\nu \geq \nu(\epsilon, \delta)$,

$$\sup_{\beta \in [0, 1-\delta]} \quad |\bar{q}_\beta\left(Z^\nu(c_0, c)\right) - \bar{q}_\beta\left(Z(c_0, c)\right)| \quad \leq \quad \frac{\epsilon}{2(1-\delta)}. \quad \text{(II.23)}$$

Then,

$$\begin{aligned}
\left| \int_0^{1-\delta} \max\left\{0, \bar{q}_\beta(Z^\nu(c_0, c))\right\} d\beta - \int_0^{1-\delta} \max\left\{0, \bar{q}_\beta(Z(c_0, c))\right\} d\beta \right| & \\
&\leq \int_0^{1-\delta} \left| \bar{q}_\beta(Z^\nu(c_0, c)) - \bar{q}_\beta(Z(c_0, c)) \right| d\beta \\
&\leq \int_0^{1-\delta} \frac{\epsilon}{2(1-\delta)} d\beta \\
&\leq \frac{\epsilon}{2} \qquad \text{(II.24)}
\end{aligned}$$

for all $\nu \geq \nu(\epsilon, \delta)$. Following an argument similar to that in Lemma II.1, we find that

$$\begin{aligned}
\int_{1-\delta}^1 \max\{0, \bar{q}_\beta(Z(c_0, c))\} d\beta \quad \leq \quad & \delta^{1/2}\sigma(Z(c_0, c)) \\
& + \max\left\{0, \delta E[Z(c_0, c)]\delta^{1/2}\sigma(Z(c_0, c))\right\}. \quad \text{(II.25)}
\end{aligned}$$

Moreover, the reasoning that leads to (II.21) also gives

$$\Big|E[Z(c_0,c)]\Big| \leq M + |c_0| + \|c\| mM. \tag{II.26}$$

These facts show that there exists a positive constant $\tilde{M} < \infty$ (which depends on c_0 and c) such that $|E[Z(c_0,c)]|, \sigma(Z(c_0,c)) \leq \tilde{M}$. Hence, from (II.25), we find that

$$\int_{1-\delta}^{1} \max\Big\{0, \bar{q}_\beta(Z(c_0,c))\Big\} d\beta \leq 3\tilde{M}\delta^{1/2}. \tag{II.27}$$

Let $\epsilon < 12\tilde{M}$ and $\delta_\epsilon = (\epsilon/(12\tilde{M}))^2$. Then, $3\tilde{M}\delta_\epsilon^{1/2} = \epsilon/4$ and

$$\int_{1-\delta_\epsilon}^{1} \max\Big\{0, \bar{q}_\beta(Z(c_0,c))\Big\} d\beta \quad \leq \quad \frac{\epsilon}{4}. \tag{II.28}$$

An identical result holds for $Z^\nu(c_0,c)$. Let $\bar{q}_\beta(Z^\nu(c_0,c))_+ = \max\{0, \bar{q}_\beta(Z^\nu(c_0,c))\}$ and $\bar{q}_\beta(Z(c_0,c))_+ = \max\{0, \bar{q}_\beta(Z(c_0,c))\}$. Consequently, for all $\nu \geq \nu(\epsilon, \delta_\epsilon)$,

$$\begin{aligned}
&\left|\int_0^1 \bar{q}_\beta(Z^\nu(c_0,c))_+ d\beta - \int_0^1 \bar{q}_\beta(Z(c_0,c))_+ d\beta\right| \\
&\qquad \leq \left|\int_0^{1-\delta_\epsilon} \bar{q}_\beta(Z^\nu(c_0,c))_+ d\beta - \int_0^{1-\delta_\epsilon} \bar{q}_\beta(Z(c_0,c))_+ d\beta\right| \\
&\qquad\qquad + \int_{1-\delta_\epsilon}^1 \bar{q}_\beta(Z^\nu(c_0,c))_+ d\beta + \int_{1-\delta_\epsilon}^1 \bar{q}_\beta(Z^\nu(c_0,c))_+ d\beta \\
&\qquad \leq \frac{\epsilon}{2} + \frac{\epsilon}{4} + \frac{\epsilon}{4} \\
&\qquad \leq \epsilon.
\end{aligned}$$

The fact that $E[Z^\nu(c_0,c)] \to E[Z(c_0,c)] < \infty$ and the assumption that (c_0,c) is arbitrary, imply that $\bar{\mathcal{E}}_\alpha(Z^\nu(\cdot,\cdot)) \to \bar{\mathcal{E}}_\alpha(Z(\cdot,\cdot))$ pointwise on $\mathbb{R}^{m+1}$. Lemma II.1 and the above moment assumptions imply that $\bar{\mathcal{E}}_\alpha(Z^\nu(\cdot,\cdot))$ and $\bar{\mathcal{E}}_\alpha(Z(\cdot,\cdot))$ are finite-valued functions. They are also convex, which follows directly from the convexity of $\bar{\mathcal{E}}_\alpha$ on $\mathcal{L}^2$ and the affine form of Z^ν and Z as functions of c_0 and c. Consequently, by Theorem 7.17 in Rockafellar and Wets (1998), $\bar{\mathcal{E}}_\alpha(Z^\nu(\cdot,\cdot))$ epiconverges to $\bar{\mathcal{E}}_\alpha(Z(\cdot,\cdot))$. The result then follows from Theorem 7.31 in Rockafellar and Wets (1998).

□

When the Approximate Superquantile Regression Problem SqR^ν is constructed using an independent identically distributed sample of size ν from the distribution of (X, Y), we obtain the following corollary which follows from the properties of the empirical distribution.

Corollary II.1. *Suppose that the basis function h is continuous except possibly on a subset $S \subset I\!R^n$ with $P\{X \in S\} = 0$ and that $H_i, Y \in \mathcal{L}^2$, $i = 1, 2, ..., m$. Moreover, let (X^ν, Y^ν) be distributed according to the empirical distribution generated by an independent and identically distributed sample of size ν from the distribution of (X, Y). Then, the conclusion of Theorem II.3 holds.*

We next examine the rate of convergence of regression vectors obtained from the approximate problem SqR^ν to those of SqR corresponding to the "true" distribution. It appears difficult to obtain asymptotic distribution theory for superquantile regression without additional assumptions, which among other consequences should ensure unique optimal solutions of SqR. We prefer another route that leads to a rate of convergence result under mild assumptions.

Quantification of the stability of the set of optimal solutions of an optimization problem under perturbations depends on a "growth condition" of the problem, which is difficult to quantify for SqR; see Section 7J in Rockafellar and Wets (1998). Consequently, we focus on the better behaved ϵ-regression vectors of SqR defined for $\epsilon > 0$ as

$$\bar{\mathcal{C}}_\epsilon := \left\{ (c_{0,\epsilon}, c_\epsilon) \in I\!R^{m+1} \,\middle|\, \bar{\mathcal{E}}_\alpha(Z(c_{0,\epsilon}, c_\epsilon)) \leq \min_{c_0 \in R, c \in R^m} \bar{\mathcal{E}}_\alpha(Z(c_0, c)) + \epsilon \right\},$$

with an analogous definition of the ϵ-regression vectors of SqR^ν denoted by $\bar{\mathcal{C}}_\epsilon^\nu$. The rate with which $\bar{\mathcal{C}}_\epsilon^\nu$ tends to $\bar{\mathcal{C}}_\epsilon$ depends, naturally, on the rate with which (X^ν, Y^ν), underlying SqR^ν, tends to (X, Y) of SqR in some sense. Before we make a precise statement, we introduce a convenient notion of distances between any two nonempty sets $A, B \subset I\!R^{m+1}$. For $\rho \geq 0$, let

$$\hat{dI}_\rho(A, B) := \inf\{\eta \geq 0 | A \cap \rho I\!B \subset B + \eta I\!B, B \cap \rho I\!B \subset A + \eta I\!B\},$$

where $I\!B$ is the Euclidean ball in $I\!R^{m+1}$ with unit radius and center at the origin. Roughly, $d\hat{I}\!\!I_\rho(A,B)$ is the smallest amount the sets need to be "enlarged" to ensure they contain the other one, with an exclusive focus on points no further from the origin than ρ. This restriction facilitates the treatment of unbounded sets.

As we see next, the rate of convergence is directly related to the rate with which the random vector

$$\Delta^\nu := (H^\nu - H, Y^\nu - Y),$$

describing the approximation error, tends to zero.

Theorem II.4. *(Rate of convergence of regression vector) Suppose that (X^ν, Y^ν), $\nu = 1, 2, ...,$ and (X, Y) are $n+1$-dimensional random vectors generating SqR^ν and SqR, respectively. Moreover, let $H_i, Y \in \mathcal{L}^2$, $\sup_\nu E[(H_i^\nu)^2] < \infty$, $i = 1, 2, ..., m$, and $\sup_\nu E[(Y^\nu)^2] < \infty$. Let $\rho_0 > 0$ be such that $\rho_0 I\!B \cap \bar{\mathcal{C}} \neq \emptyset$ and $\rho_0 I\!B \cap \bar{\mathcal{C}}^\nu \neq \emptyset$.*

Then, for $\rho > \rho_0$, there exist positive constants k_1, k_2, and k_3 (dependent on ρ) such that for any $\epsilon > 0$ and $\nu = 1, 2, ...,$

$$d\hat{I}\!\!I_\rho(\bar{\mathcal{C}}_\epsilon^\nu, \bar{\mathcal{C}}_\epsilon) \leq \left(1 + \frac{4\rho}{\epsilon}\right)\left[E[\|\Delta^\nu\|]\left(k_1 \max\left\{0, \log\left(\frac{1}{E[\|\Delta^\nu\|]}\right)\right\} + k_2\right) + k_3\|E[\Delta^\nu]\|\right]$$

whenever $E[\|\Delta^\nu\|] > 0$ and $d\hat{I}\!\!I_\rho(\bar{\mathcal{C}}_\epsilon^\nu, \bar{\mathcal{C}}_\epsilon) = 0$ otherwise.

Proof.

By Theorem 3(a) of Rockafellar and Royset (2014b), for $\beta \in [0, 1)$,

$$\left|\bar{q}_\beta(Z^\nu(c_0, c)) - \bar{q}_\beta(Z(c_0, c))\right| \;\leq\; \frac{1}{1-\beta} E[|Z^\nu(c_0, c) - Z(c_0, c)|],$$

and since

$$\frac{1}{1-\beta} E[|Z^\nu(c_0, c) - Z(c_0, c)|] \;=\; \frac{1}{1-\beta} E[|\langle \tilde{c}, \Delta^\nu \rangle|],$$

we get that

$$\begin{aligned} \left|\bar{q}_\beta(Z^\nu(c_0, c)) - \bar{q}_\beta(Z(c_0, c))\right| \;&\leq\; \frac{1}{1-\beta} E[|\langle \tilde{c}, \Delta^\nu \rangle|] \\ &\leq\; \frac{1}{1-\beta} \|\tilde{c}\| E[\|\Delta^\nu\|], \end{aligned} \tag{II.29}$$

where $\tilde{c} = (-c, 1)$. Then, for $\delta \in (0, 1)$,

$$\begin{aligned}\left| \int_0^{1-\delta} \max\{0, \bar{q}_\beta(Z^\nu(c_0, c))\} d\beta - \int_0^{1-\delta} \max\{0, \bar{q}_\beta(Z(c_0, c))\} d\beta \right| \\ &\leq \int_0^{1-\delta} |\bar{q}_\beta(Z^\nu(c_0, c)) - \bar{q}_\beta(Z(c_0, c))| \, d\beta \\ &\leq \|\tilde{c}\| E[\|\Delta^\nu\|] \int_0^{1-\delta} \frac{1}{1-\beta} d\beta \\ &\leq -\|\tilde{c}\| E[\|\Delta^\nu\|] \log \delta. \end{aligned} \tag{II.30}$$

Let $\rho > \rho_0$ and M be an upper bound on first and second moments of $|H_i|$, $|H_i^\nu|$, $|Y|$, and $|Y^\nu|$ as in the proof of Theorem II.3. Since $|\langle c, H\rangle| \leq \|c\| \sum_{i=1}^m |H_i|$ and $\langle c, H\rangle^2 \leq \|c\|^2 \sum_{i=1}^m (H_i)^2$, we find that $E[|\langle c, H\rangle|] \leq \|c\| mM$ and $E[\langle c, H\rangle^2] \leq \|c\|^2 mM$. Consequently,

$$\begin{aligned} E[(Y - c_0 - \langle c, H\rangle)^2] &\leq E[(Y - c_0)^2] + 2|E[(Y - c_0)\langle c, H\rangle]| + E[\langle c, H\rangle^2] \\ &\leq M + 2(\|c\| m^{1/2} M + (M + |c_0|)\|c\| mM) + \|c\|^2 mM. \end{aligned} \tag{II.31}$$

Then, for $\|(c_0, c)\| \leq \rho$, it follows by (II.26) that

$$|E[Z(c_0, c)]| \leq M + \rho + \rho mM$$

and by (II.31) that

$$\sigma(Z(c_0, c)) \leq \left(M + 2\left(\rho m^{1/2} M + (M + \rho)\rho mM\right) + \rho^2 mM \right)^{1/2},$$

with identical bounds for $|E[Z^\nu(c_0, c)]|$ and $\sigma(Z^\nu(c_0, c))$. Let M_ρ be the larger of the two previous right-hand sides.

By (II.25), analogously to (II.27), we have that for $\|(c_0, c)\| \leq \rho$,

$$\int_{1-\delta}^1 \max\{0, \bar{q}_\beta(Z(c_0, c))\} d\beta \leq 3 M_\rho \delta^{1/2} \tag{II.32}$$

and similarly with $Z(c_0, c)$ replaced by $Z^\nu(c_0, c)$.

We also find that for $\|(c_0, c)\| \leq \rho$,

$$\begin{aligned}\left|E[Z^\nu(c_0,c)] - E[Z(c_0,c)]\right| &= \left|\langle \tilde{c}, E[\Delta^\nu]\rangle\right| \\ &\leq \|\tilde{c}\|\|E[\Delta^\nu]\| \\ &\leq (1+\rho)\|E[\Delta^\nu]\|. \end{aligned} \tag{II.33}$$

Then, collecting the results of (II.30), (II.32), and (II.33), and for $\|(c_0, c)\| \leq \rho$, we obtain

$$\begin{aligned}&\left|\bar{\mathcal{E}}_\alpha(Z^\nu(c_0,c)) - \bar{\mathcal{E}}_\alpha(Z(c_0,c))\right| \\ &\qquad \leq \left|\int_0^1 \max\{0, \bar{q}_\beta(Z^\nu(c_0,c))\}d\beta - \int_0^1 \max\{0, \bar{q}_\beta(Z(c_0,c))\}d\beta\right| \\ &\qquad\qquad + \left|E[Z^\nu(c_0,c)] - E[Z(c_0,c)]\right| \\ &\qquad \leq \left|\int_0^{1-\delta} \max\{0, \bar{q}_\beta(Z^\nu(c_0,c))\}d\beta - \int_0^{1-\delta} \max\{0, \bar{q}_\beta(Z(c_0,c))\}d\beta\right| \\ &\qquad\qquad + \int_{1-\delta}^1 \max\{0, \bar{q}_\beta(Z^\nu(c_0,c))\}d\beta + \int_{1-\delta}^1 \max\{0, \bar{q}_\beta(Z(c_0,c))\}d\beta \\ &\qquad\qquad + \left|E[Z^\nu(c_0,c)] - E[Z(c_0,c)]\right| \\ &\qquad \leq -(1+\rho)E[\|\Delta^\nu\|]\log\delta + 6M_\rho\delta^{1/2} + (1+\rho)\|E[\Delta^\nu]\|. \end{aligned} \tag{II.34}$$

We next determine the choice of $\delta \in (0,1)$ that minimizes the previous bound and consider two cases. First, if

$$0 < k_\rho \left(E[\|\Delta^\nu\|]\right)^2 < 1,$$

with

$$k_\rho := \left(\frac{2(1+\rho)}{6M_\rho}\right)^2,$$

then differentiation gives that the bound is minimized with $\delta = k_\rho(E[\|\Delta^\nu\|])^2$. Second, if

$$k_\rho \left(E[\|\Delta^\nu\|]\right)^2 \geq 1,$$

then

$$M_\rho \leq \frac{4(1+\rho)E[\|\Delta^\nu\|]}{6}$$

and the bound

$$-(1+\rho)E[\|\Delta^\nu\|]\log\delta + 6M_\rho\delta^{1/2} + (1+\rho)\|E[\Delta^\nu]\|$$
$$\leq -(1+\rho)E[\|\Delta^\nu\|]\log\delta + 4(1+\rho)E[\|\Delta^\nu\|]\delta^{1/2} + (1+\rho)\|E[\Delta^\nu]\|,$$

for any $\delta \in (0,1)$. Consequently, combining the two cases, there exist constants k_1, k_2, and k_3 (which depend on ρ), such that for $\|(c_0,c)\| \leq \rho$,

$$\left|\bar{\mathcal{E}}_\alpha(Z^\nu(c_0,c)) - \bar{\mathcal{E}}_\alpha(Z(c_0,c))\right|$$
$$\leq k_1 E[\|\Delta^\nu\|]\max\left\{0, \log\left(\frac{1}{E[\|\Delta^\nu\|]}\right)\right\} + k_2 E[\|\Delta^\nu\|] + k_3\|E[\Delta^\nu]\|$$
$$\leq E[\|\Delta^\nu\|]\left(k_1\max\left\{0,\log\left(\frac{1}{E[\|\Delta^\nu\|]}\right)\right\} + k_2\right) + k_3\|E[\Delta^\nu]\|.$$

Direct application of Example 7.62 and Theorem 7.69 of Rockafellar and Wets (1998) then yields the conclusion for $E[\|\Delta^\nu\|] > 0$, where the additional coefficient $(1+4\rho/\epsilon)$ originates in that theorem. Finally, if $E[\|\Delta^\nu\|] = 0$, then, in view of (II.29) and the fact that this implies that $\|E[\Delta^\nu]\| = 0$, we find that for $\|(c_0,c)\| \leq \rho$,

$$\left|\bar{\mathcal{E}}_\alpha(Z^\nu(c_0,c)) - \bar{\mathcal{E}}_\alpha(Z(c_0,c))\right| = 0.$$

The final conclusion then follows by again invoking Example 7.62 and Theorem 7.69 of Rockafellar and Wets (1998).

□

Theorem II.4 shows that the distance between $\bar{\mathcal{C}}_\epsilon^\nu$ and $\bar{\mathcal{C}}_\epsilon$ is almost proportional to $E[\|\Delta^\nu\|]$, but with a minor correction by a logarithmic term. If the approximation (X^ν, Y^ν) is caused by measurement errors of magnitude $1/\nu$, i.e., the absolute value of each component of $(X^\nu - X, Y^\nu - Y)$ is no greater than $1/\nu$ almost surely, then $E[\|\Delta^\nu\|] \leq \sqrt{m+1}/\nu$ and the expressions can be simplified. For $\xi > 0$, $\log x \leq x^\xi$ for sufficiently large $x \in I\!R$. Consequently, for any $\xi \in (0,1)$ and sufficiently large ν,

$$\hat{dI}_\rho(\bar{\mathcal{C}}_\epsilon^\nu, \bar{\mathcal{C}}_\epsilon) \leq \left(1 + \frac{4\rho}{\epsilon}\right)\frac{k}{\nu^{1-\xi}},$$

where $k > 0$ can be determined from $k_1,\ k_2,\ k_3$, and m. That is, the Euclidean distance between an ϵ-regression vector of SqR^ν to one of SqR is $O(\nu^{\xi-1})$ for $\xi \in (0,1)$ arbitrarily close to zero.

2. Superquantile Tracking

We next turn to the situation where the primary concern is with the conditional loss $Y(x)$ given that the explanatory random variable takes on specific values, $X = x$. We seek to estimate $\bar{q}_\alpha(Y(x))$ for $x \in I\!R^n$, or a subset thereof, with the goal of eventually minimizing, at least approximately, $\bar{q}_\alpha(Y(x))$ by a judicious choice of x. Of course, with incomplete knowledge about the distributions of $Y(x)$ this is a difficult task that can be achieved only approximately. For example, there is no guarantee that a regression function $f = \bar{c}_0 + \langle \bar{c}, h(\cdot)\rangle$, with $(\bar{c}_0, \bar{c}) \in \bar{\mathcal{C}}$ obtained by solving SqR using $\alpha \in (0,1)$, tracks $\bar{q}_\alpha(Y(x))$, i.e., $f(x) = \bar{q}_\alpha(Y(x))$ for all $x \in I\!R^n$. The hope of such "exact" superquantile tracking becomes even less realistic when SqR must be replaced by an approximation SqR^ν as typically required in practice. However, "local" superquantile tracking is possible, at least approximately, as stated in the next proposition. Moreover, tracking is achieved under certain model assumptions. For example, if we have that $Y = \bar{c}_0 + \langle \bar{c}, X\rangle + \epsilon$, for some $\bar{c}_0 \in I\!R$, $\bar{c} \in I\!R^n$, and where ϵ is independent of X, then superquantile tracking is guaranteed; see Theorem 5.1 in Rockafellar and Royset (2014a).

Here we consider the situation where there is a sample of $Y(x)$ for some values of x, but this sample is not large enough to allow pointwise estimation of $\bar{q}_\alpha(Y(x))$ for every x of interest. There may even be no x for which there are multiple sample points of $Y(x)$. Concentrating on a particular $\hat{x} \in I\!R^n$, we hope to estimate $\bar{q}_\alpha(Y(\hat{x}))$ by using samples from $Y(x)$ for x near $\hat{x}$, weighted appropriately. The weights should be nonnegative, sum to one, and can be thought of as an artificially constructed probability distribution associated with the sample. Specifically, suppose that $x_i,\ i = 1, ..., \nu$, are the points where the sample is observed and $y_i,\ i = 1, ..., \nu$, are the corresponding realizations of $Y(x_i)$. When estimating a superquantile at $\hat{x}$, we put

more "trust" on sample points taken near $\hat{x}$ and consequently the weight of (x_i, y_i) may be inversely proportional to $\|x_i - \hat{x}\|$, with an appropriate adjustment if $\hat{x}$ coincides with an x_i.

A justification for the approach follows directly from Theorem II.3 through the next proposition.

Proposition II.5. *Suppose that the assumptions of Theorem II.3 hold and that the probability distribution of (X, Y) is degenerate at $\hat{x} \in \mathbb{R}^{n+1}$ in the sense that $P\{(X, Y) \leq (x, y)\} = \varphi(y)$, for all $y \in \mathbb{R}$ and $x \geq \hat{x}$, where $\varphi(y) = P\{Y(\hat{x}) \leq y\}$, and $P\{(X, Y) \leq (x, y)\} = 0$ otherwise.*

If $\{(\bar{c}_0^\nu, \bar{c}^\nu)\}_{\nu=1}^\infty$ is a sequence of optimal solutions of SqR^ν, with $\alpha \in (0, 1)$, then along every convergent subsequence we have that $\bar{c}_0^\nu + \langle \bar{c}^\nu, h(\hat{x})\rangle$ tends to $\bar{q}_\alpha(Y(\hat{x}))$.

Proof.

For the given degenerate distribution of (X, Y), $c_0 + \langle c, h(X)\rangle = c_0 + \langle c, h(\hat{x})\rangle$ almost surely. Consequently, SqR reduces to the error minimization problem of Theorem II.1 and $\bar{c}_0 + \langle \bar{c}, h(\hat{x})\rangle = \bar{q}_\alpha(Y(\hat{x}))$ for every $(\bar{c}_0, \bar{c}) \in \bar{\mathcal{C}}$. The conclusion then follows from Theorem II.3.

□

Suppose that the weights of (x_i, y_i), $i = 1, 2, ..., \nu$, in the above construction are chosen to approximate the degenerate distribution of Proposition II.5, for example by setting them inversely proportional to $\|x_i - \hat{x}\|$. Then, in view of Proposition II.5, a solution of SqR^ν, constructed using those weights as an artificial probability distribution for (X^ν, Y^ν), leads to an approximation of the considered superquantile at $\hat{x}$. Of course, this procedure can be repeated for different points $\hat{x}$ to generate a "global" assessment of $\bar{q}_\alpha(Y(x))$ as a function of x and eventually facilitate optimization over x. Moreover, the process can be repeated with new or augmented sample points in a straightforward manner. In a situation where a sample is not fully randomly generated but x-points are determined by an analyst, the approach may even motivate scattering those points near a point of interest $\hat{x}$ instead of concentrating them all at $\hat{x}$ exactly. The former approach certainly results in a better "global" understanding

of a superquantile as a function of x, but may prove to be a more economical route to estimate a superquantile at $\hat{x}$ too. We examine this situation numerically in Chapter IV.

C. VALIDATION ANALYSIS

Regression modeling must be associated with means of assessing the goodness of fit of a computed regression vector. The process of validating a regression fit is important as it allows us to decide whether the obtained numerical results quantify how well the model explains and predicts future outcomes. A commonly used measure that allows such assessment is the coefficient of determination. In least squares regression, this coefficient, also known as R-squared, is defined as

$$R^2 = 1 - \frac{SS_{\text{Res}}}{SS_{\text{T}}},$$

where SS_{Res} denotes the residual sum of squares and SS_{T} the total sum of squares. While R^2 cannot be relied on exclusively, it provides an indication of the goodness of fit that is easily extended to the present context of superquantile regression. In our notation,

$$R^2 = 1 - \frac{E[Z(c_0, c)^2]}{\sigma^2(Y)}, \tag{II.35}$$

and similarly when passing to an approximate random vector (X^ν, Y^ν). From Example 1' in Rockafellar and Uryasev (2013), we know that the numerator in (II.35) is a measure of error applied to $Z(c_0, c)$ and that its denominator corresponds to the measure of deviation $\sigma^2(\cdot)$. Moreover, the minimization of that error of $Z(c_0, c)$ results in the least squares regression vector. According to Rockafellar and Uryasev (2013), these measures of error and deviation are in correspondence and belong to a family of risk quadrangles that yields the expectation as its statistic. Therefore we could write the formula for R^2 as follows

$$R^2 = 1 - \frac{\mathcal{E}(Z(c_0, c))}{\mathcal{D}(Y)}. \tag{II.36}$$

This observation motivates the following definition of coefficient of determination applied to quantile regression.

Definition II.1. In quantile regression, the coefficient of determination of a regression vector $(c_0, c) \in I\!R^{m+1}$ is given by

$$\begin{aligned} R^2_\alpha(c_0, c) &:= 1 - \frac{\mathcal{E}_\alpha(Z(c_0, c))}{\mathcal{D}_\alpha(Y)} \\ &= 1 - \frac{E\left[\frac{\alpha}{1-\alpha} Z(c_0, c)_+ + Z(c_0, c)_-\right]}{\bar{q}_\alpha(Y) - E[Y]}, \end{aligned} \tag{II.37}$$

where $Z(c_0, c)_+ = \max\{0, Z(c_0, c)\}$ and $Z(c_0, c)_- = \max\{0, -Z(c_0, c)\}$.

In least squares regression, the coefficient of determination is a value expressed between zero and one, which leads us to the following proposition.

Proposition II.6. *For a regression vector $(c_0, c) \in I\!R^{m+1}$ and $\alpha \in (0, 1)$, one has that*

$$0 \;\leq\; R^2_\alpha(c_0, c) \;\leq\; 1. \tag{II.38}$$

Proof.

By the definition of coefficient of determination in quantile regression and of quantile error and deviation measures, in the sense of Rockafellar and Uryasev (2013), we have that

$$\begin{aligned} R^2_\alpha(c_0, c) &= 1 - \frac{\mathcal{E}_\alpha(Z(c_0, c))}{\mathcal{D}_\alpha(Y)} \\ &= 1 - \frac{\mathcal{E}_\alpha(Y - c_0 - \langle c, h(X)\rangle)}{\min\limits_{\xi \in R} \;\; \mathcal{E}_\alpha(Y - \xi)} \\ &= 1 - \frac{\mathcal{E}_\alpha(Y - c_0 - \langle c, h(X)\rangle)}{\mathcal{E}_\alpha(Y - \xi^*)}, \end{aligned} \tag{II.39}$$

where ξ^* is an optimal solution to $\min_{\xi \in R} \;\; \mathcal{E}_\alpha(Y - \xi)$. Both quantile error and deviation measures are nonnegative quantities, which proves the upper bound. Since the regression vector $(c_0, c) \in I\!R^{m+1}$ is obtained by

$$\min_{(c_0, c) \in R^{m+1}} \;\; \mathcal{E}_\alpha\left(Y - c_0 - \langle c, h(X)\rangle\right),$$

we guarantee that

$$\mathcal{E}_\alpha(Y - c_0 - \langle c, h(X)\rangle) \;\leq\; \mathcal{E}_\alpha(Y - \xi^*),$$

in equation (II.39), which gives $R^2_\alpha(c_0, c) \geq 0$.

□

Applying the same idea to superquantile regression, we obtain the following definition of the coefficient of determination.

Definition II.2. In superquantile regression, the coefficient of determination of a regression vector $(c_0, c) \in I\!R^{m+1}$ is given by

$$\begin{aligned} \bar{R}^2_\alpha(c_0, c) &:= 1 - \frac{\bar{\mathcal{E}}_\alpha(Z(c_0, c))}{\bar{\mathcal{D}}_\alpha(Y)} \\ &= 1 - \frac{\frac{1}{1-\alpha}\int_0^1 \max\{0, \bar{q}_\beta(Z(c_0, c))\}\, d\beta - E\left[Z(c_0, c)\right]}{\frac{1}{1-\alpha}\int_\alpha^1 \bar{q}_\beta(Y) d\beta - E[Y]}. \end{aligned} \tag{II.40}$$

In fact, a similar definition can be formulated for any generalized regression consisting of minimizing an error of Z_f.

The bounds on the coefficients of determination for least squares and quantile regressions, can also be applied to the coefficient of determination in the superquantile regression case, using the same arguments as in the previous proof.

Proposition II.7. *For a regression vector $(c_0, c) \in I\!R^{m+1}$ and $\alpha \in (0, 1)$, one has that*

$$0 \;\leq\; \bar{R}^2_\alpha(c_0, c) \;\leq\; 1. \tag{II.41}$$

Proof.

By the definition of coefficient of determination in superquantile regression and of superquantile error and deviation measures, in the sense of Rockafellar and Uryasev

(2013), we have that

$$\begin{aligned}\bar{R}^2_\alpha(c_0, c) &= 1 - \frac{\bar{\mathcal{E}}_\alpha(Z(c_0, c))}{\bar{\mathcal{D}}_\alpha(Y)} \\ &= 1 - \frac{\bar{\mathcal{E}}_\alpha(Y - c_0 - \langle c, h(X)\rangle)}{\min\limits_{\xi \in R} \ \bar{\mathcal{E}}_\alpha(Y - \xi)} \\ &= 1 - \frac{\bar{\mathcal{E}}_\alpha(Y - c_0 - \langle c, h(X)\rangle)}{\bar{\mathcal{E}}_\alpha(Y - \xi^*)}, \end{aligned} \tag{II.42}$$

where ξ^* is an optimal solution to $\min_{\xi \in R} \ \bar{\mathcal{E}}_\alpha(Y - \xi)$. Using the same arguments as in the proof of Proposition II.6, we arrive at the conclusion.

□

As in the classical case, higher values of $\bar{R}^2_\alpha$ are better, at least in some sense. Indeed, SqR aims to minimize the error of $Z(c_0, c)$ by wisely selecting the regression vector (c_0, c) and thereby also maximizes $\bar{R}^2_\alpha$,

$$\operatorname*{argmin}_{c_0, c} \bar{\mathcal{E}}_\alpha \left(Y - [c_0 + \langle c, h(X)\rangle]\right) \Leftrightarrow \operatorname*{argmax}_{c_0, c} \bar{R}^2_\alpha(c_0, c). \tag{II.43}$$

The error is "normalized" with the overall "nonconstancy" in Y as measured by its measure of deviation to more easily allow for comparison of coefficients of determination across data sets.

However it is possible to obtain large coefficients of determination by adding explanatory terms to a regression model, i.e., increasing m, but without necessarily achieving a more useful model. Hence, it is usual in least squares regression to also evaluate an adjusted coefficient of determination that penalizes any term added to the model that does not reduce variability substantially. This quantity only increases if a new term reduces $SS_{\text{Res}}/(\nu - m - 1)$ as seen by the definition

$$R^2_{\text{Adj}} = 1 - \frac{SS_{\text{Res}}/(\nu - m - 1)}{SS_{\text{T}}/(\nu - 1)}, \tag{II.44}$$

where ν is the number of observations. Naturally, then, we define an adjusted coefficient of determination for quantile and superquantile regressions similarly in the case where the distribution of (X, Y) has a finite support of cardinality ν.

Definition II.3. In quantile regression, the adjusted coefficient of determination of a regression vector $(c_0, c) \in I\!R^{m+1}$ is given by

$$R^2_{\alpha,\text{Adj}}(c_0, c) := 1 - \frac{\mathcal{E}_\alpha(Z(c_0, c))/(\nu - m - 1)}{\mathcal{D}_\alpha(Y)/(\nu - 1)}. \tag{II.45}$$

Definition II.4. In superquantile regression, the adjusted coefficient of determination of a regression vector $(c_0, c) \in I\!R^{m+1}$ is given by

$$\bar{R}^2_{\alpha,\text{Adj}}(c_0, c) := 1 - \frac{\bar{\mathcal{E}}_\alpha(Z(c_0, c))/(\nu - m - 1)}{\bar{\mathcal{D}}_\alpha(Y)/(\nu - 1)}. \tag{II.46}$$

Again, similar expressions are available for other generalized regression techniques.

When performing least squares regression analysis, we have other commonly used validation methods. These include computing the Cook's distance for each observation used in the model, which provides an estimate on how an observation influences the obtained regression fit. This distance allows the decision maker to easily understand which observation might be considered an outlier and which should be checked for validity. In least squares regression, the Cook's distance for observation i is defined as

$$D_i = \frac{\left(f(X) - f^{(i)}(X)\right)^2}{m\text{MSE}} = \frac{\left(f(X) - f^{(i)}(X)\right)^2}{m\left(f(X) - Y\right)^2}, \tag{II.47}$$

where $f^{(i)}(\cdot)$ represents the fitted regression function without observation i, and MSE denotes the mean square error of the regression model. Using the measure of error that is corresponding to the expectation as the statistic, and similar to the approach in (II.36), we could write the formula for the Cook's distance in (II.47) for observation i as follows

$$D_i := \frac{\mathcal{E}(f(X) - f^{(i)}(X))}{m\ \mathcal{E}(f(X) - Y)}. \tag{II.48}$$

With this in mind, we next define Cook's distances applied to quantile and superquantile regressions.

Definition II.5. In quantile regression, the Cook's distance estimates for a regression vector $(c_0, c) \in I\!R^{m+1}$ is given by

$$\begin{aligned} D_{i,\alpha}(c_0, c) &:= \frac{\mathcal{E}_\alpha(f(X) - f^{(i)}(X))}{m\ \mathcal{E}_\alpha(f(X) - Y)} \\ &= \frac{E\left[\frac{\alpha}{1-\alpha}\{f(X) - f^{(i)}(X)\}_+ + \{f(X) - f^{(i)}(X)\}_-\right]}{mE\left[\frac{\alpha}{1-\alpha}\{f(X) - Y\}_+ + \{f(X) - Y\}_-\right]}, \end{aligned} \tag{II.49}$$

where $Y_+ = \max\{0, Y\}$ and $Y_- = \max\{0, -Y\}$.

Definition II.6. In superquantile regression, the Cook's distance estimates for a regression vector $(c_0, c) \in I\!R^{m+1}$ is given by

$$\begin{aligned} \bar{D}_{i,\alpha}(c_0, c) &:= \frac{\bar{\mathcal{E}}_\alpha(f(X) - f^{(i)}(X))}{m\ \bar{\mathcal{E}}_\alpha(-Z(c_0, c))} \\ &= \frac{\frac{1}{1-\alpha}\int_0^1 \max\{0, \bar{q}_\alpha(f(X) - f^{(i)}(X))\}d\beta - E[f(X) - f^{(i)}(X)]}{m\frac{1}{1-\alpha}\int_0^1 \max\{0, \bar{q}_\alpha(-Z(c_0, c))\}d\beta + E[Z(c_0, c)]}. \end{aligned} \tag{II.50}$$

As shown in Section II.B, we only use one assumption when building our superquantile regression problem, finite second moments for the random variables. This generalization allows the regression problem to be applied in many situations, but makes validating the obtained model a harder process.

For the scope of the dissertation we do not develop other model validation techniques since we discard many of the commonly used model assumptions, such as normality, or homoscedasticity, that are usually requirements for such assessment tests. However, we recall that cross-validation is a tool to take into account for validating the regression model, especially for larger sample sizes ν. Obviously, when the sample size is small and we choose a high probability level α, subdividing the sample into training and testing data sets is not a wise decision.

In the next chapter, we develop computational methods that allow us to implement these theoretical results.

III. COMPUTATIONAL METHODS

In this chapter, we develop computational methods that allow us to solve the superquantile regression problems of Section II.B. This computational task consists of solving the convex optimization problem SqR, or in practice the approximate problem SqR^ν due to incomplete distributional information.

In the next two sections, we describe convenient means for solving the superquantile regression problems when (X^ν, Y^ν) has a discrete joint distribution with ν possible realizations. Regardless of the distribution of (X^ν, Y^ν), a reformulation of the approximate problem SqR^ν in terms of the deviation measure $\bar{\mathcal{D}}_\alpha$ is beneficial. In view of Proposition II.3, the task of determining a regression vector $(\bar{c}_0^\nu, \bar{c}^\nu)$ reduces to that of minimizing $\bar{\mathcal{D}}_\alpha(Z_0^\nu(\cdot))$, obtaining $\bar{c}^\nu$ as an optimal solution, and then setting $\bar{c}_0^\nu = \bar{q}_\alpha(Z_0^\nu(\bar{c}^\nu))$.

Since it is straightforward to compute every superquantile of a random variable Y with a discrete probability distribution, as follows

$$\bar{q}_\alpha(Y) = \begin{cases} \sum_{j=1}^{\nu} p_j y_j & \text{if } \alpha = 0, \\ \frac{1}{1-\alpha}\left[\left(\sum_{j=1}^{i} p_i - \alpha\right) y_i + \sum_{j=i+1}^{\nu} p_j y_j\right] & \text{if } \sum_{j=1}^{i-1} p_j < \alpha \leq \sum_{j=1}^{i} p_j < 1, \\ y_\nu & \text{if } \alpha > 1 - p_\nu, \end{cases}$$

with p_j being the corresponding probabilities of the realizations y_j of Y, which are ordered from smaller to larger, we only focus on the minimization problem, which takes the following form:

Problem:

$$DSqR^\nu: \qquad \min_{c \in R^m} \ \bar{\mathcal{D}}_\alpha\left(Z_0^\nu(c)\right) = \frac{1}{1-\alpha}\int_\alpha^1 \bar{q}_\beta(Z_0^\nu(c))d\beta - E[Z_0^\nu(c)]$$

We denote these computational methods by primal methods since we compute the regression vectors solving the original problem. The material in Section III.A is to a large extend based on our paper Rockafellar et al. (2014).

In Section III.B, we use a different approach that relies on the dualization of risk and using the theory developed in Rockafellar and Royset (2014c), we generate a computational method that we denote as the dual method.

A. PRIMAL METHODS

The next subsections describe two primal computational methods for solving $DSqR^{\nu}$. The first one solves our problem by analytical integration, while the second one utilizes numerical integration techniques.

1. Analytical Integration

At first one might get the impression that numerical integration is required for solving $DSqR^{\nu}$, but this may not actually be needed as we show next. Suppose that (X^{ν}, Y^{ν}) has a discrete distribution with support (x^j, y^j), $j = 1, 2, ..., \nu$ and probability of occurring $P\{(X^{\nu}, Y^{\nu}) = (x^j, y^j)\} = 1/\nu$ for $j = 1, 2, ..., \nu$. This is the case we typically encounter in applications, where (x^j, y^j), $j = 1, 2, ..., \nu$, is the data assumed to be equally likely to occur. We then obtain significant simplifications in the approximate regression problem $DSqR^{\nu}$.

For any fixed $c \in I\!R^m$, the cumulative distribution function of $Z_0^{\nu}(c)$ is a piecewise constant function with at most ν steps. The range of the distribution function is $\{0, 1/\nu, 2/\nu, ..., 1\}$ or a subset thereof. By partitioning the integral over β in $DSqR^{\nu}$ according to this range, and accounting for the fact that the integral starts at α, we can then rewrite the optimization problem in this case as

$$\min_{c \in R^m} \quad \frac{1}{1-\alpha} \sum_{i=\nu_\alpha}^{\nu} \int_{\beta_{i-1}}^{\beta_i} \bar{q}_\beta(Z_0^{\nu}(c)) d\beta \; - \; E[Z_0^{\nu}(c)], \tag{III.1}$$

where $\nu_\alpha := \lceil \nu\alpha \rceil$, with $\lceil a \rceil$ being the smallest integer no smaller than $a \in I\!R$, $\beta_{\nu_\alpha - 1} = \alpha$, and $\beta_i = i/\nu$, for $i = \nu_\alpha, \nu_\alpha + 1, ..., \nu$.

We recall that

$$q_\alpha(Y) \in \operatorname*{argmin}_{c_0 \in R} \{c_0 + \mathcal{V}_\alpha(Y - c_0)\}$$
$$\bar{q}_\alpha(Y) = \min_{c_0 \in R} \{c_0 + \mathcal{V}_\alpha(Y - c_0)\}.$$

Consequently,

$$\begin{aligned} \bar{q}_\beta(Z_0^\nu(c)) &= \min_{U_\beta \in R} U_\beta + \frac{1}{1-\beta}\, E[\max\{Z_0^\nu(c) - U_\beta, 0\}] \\ &= q_\beta(Z_0^\nu(c)) + \frac{1}{1-\beta}\, E[\max\{Z_0^\nu(c) - q_\beta(Z_0^\nu(c)), 0\}] \end{aligned} \tag{III.2}$$

for each $\beta \in [0, 1)$.

The special piecewise-constant structure of the cumulative distribution function of $Z_0^\nu(c)$ implies that $q_\beta(Z_0^\nu(c))$ is constant as a function of β on the intervals (β_{i-1}, β_i), for every $i = \nu_\alpha, \nu_\alpha + 1, ..., \nu$. Consequently, U_β, for $\beta \in (\alpha, 1)$ in equation (III.2) can be replaced by a finite number of variables so that equation (III.1) takes the form

$$\min_{c \in R^m} \quad \frac{1}{1-\alpha} \sum_{i=\nu_\alpha}^{\nu} \int_{\beta_{i-1}}^{\beta_i} \min_{U_i \in R} \left(U_i + \frac{1}{1-\beta} E[\max\{Z_0^\nu(c) - U_i, 0\}] \right) d\beta - E[Z_0^\nu(c)].$$

The last integral simplifies further since for $\beta \in (\beta_{\nu-1}, \beta_\nu) = (1 - 1/\nu, 1)$, we have that

$$\bar{q}_\beta(Z_0^\nu(c)) = M(c) := \max_{j=1,2,...,\nu} y^j - \langle c, x^j \rangle,$$

and therefore

$$\begin{aligned} \frac{1}{1-\alpha} \int_{\beta_{\nu-1}}^{\beta_\nu} \min_{U_\nu \in R} \left(U_\nu + \frac{1}{1-\beta} E[\max\{Z_0^\nu(c) - U_\nu, 0\}] \right) d\beta \\ = \frac{1}{1-\alpha} M(c) \int_{\beta_{\nu-1}}^{\beta_\nu} d\beta = \frac{M(c)}{\nu(1-\alpha)}. \end{aligned}$$

Consequently, equation (III.1) takes the form

$$\begin{aligned} \min_{c \in R^m} \quad \frac{1}{1-\alpha} \sum_{i=\nu_\alpha}^{\nu-1} \int_{\beta_{i-1}}^{\beta_i} \min_{U_i \in R} \left(U_i + \frac{1}{1-\beta} E[\max\{Z_0^\nu(c) - U_i, 0\}] \right) d\beta \\ + \frac{M(c)}{\nu(1-\alpha)} \; - \; E[Z_0^\nu(c)]. \end{aligned}$$

The order of minimization is immaterial and we can equivalently consider

$$\min_{c\in R^m,\ U\in R^{\nu-\nu_\alpha}} \quad \frac{1}{1-\alpha}\sum_{i=\nu_\alpha}^{\nu-1}\int_{\beta_{i-1}}^{\beta_i}\left(U_i+\frac{1}{1-\beta}E[\max\{Z_0^\nu(c)-U_i,0\}]\right)d\beta \\ +\frac{M(c)}{\nu(1-\alpha)} \ - \ E[Z_0^\nu(c)],$$

where we let $U=(U_{\nu_\alpha},U_{\nu_\alpha+1},...,U_{\nu-1})\in I\!R^{\nu-\nu_\alpha}$.

In order to simplify the notation in our minimization problem, we define a_i, for $i=\nu_\alpha,\nu_\alpha+1,...,\nu-1$, as follows

$$a_i:=\int_{\beta_{i-1}}^{\beta_i}\frac{1}{1-\beta}d\beta=\log(1-\beta_{i-1})-\log(1-\beta_i).$$

Using this notation, equation (III.1) simplifies even further to

$$\min_{c\in R^m,\ U\in R^{\nu-\nu_\alpha}} \quad \frac{1}{1-\alpha}\sum_{i=\nu_\alpha}^{\nu-1}(\beta_i-\beta_{i-1})U_i+\frac{1}{1-\alpha}\sum_{i=\nu_\alpha}^{\nu-1}E[\max\{Z_0^\nu(c)-U_i,0\}]a_i \\ +\frac{M(c)}{\nu(1-\alpha)} \ - \ E[Z_0^\nu(c)].$$

By introducing another set of auxiliary variables and using the standard transcription technique for handling max-functions, we reach the linear program P_{LP}^ν that implements analytical integration.

Problem:

$$
\begin{aligned}
P^{\nu}_{\mathrm{LP}}: \quad \min_{c,u,v,w} \ & \frac{1}{1-\alpha}\sum_{i=\nu_\alpha}^{\nu-1}(\beta_i-\beta_{i-1})u_i \ + \ \frac{1}{\nu(1-\alpha)}\sum_{i=\nu_\alpha}^{\nu-1}\sum_{j=1}^{\nu} a_i v_{ij} \\
& + \ \frac{w}{\nu(1-\alpha)} \ - \ \frac{1}{\nu}\sum_{j=1}^{\nu}(y^j - \langle c, h(x^j)\rangle) \\
\text{s.t.} \quad & y^j - \langle c, h(x^j)\rangle - u_i \ \leq \ v_{ij}, \quad i = \nu_\alpha, \ldots, \nu-1 \ , j = 1, \ldots, \nu \\
& 0 \ \leq \ v_{ij}, \quad i = \nu_\alpha, \ldots, \nu-1 \ , j = 1, \ldots, \nu \\
& y^j - \langle c, h(x^j)\rangle \ \leq \ w, \quad j = 1, \ldots, \nu \\
& c \ \in \ I\!R^m \\
& u = (u_{\nu_\alpha}, \ldots, u_{\nu-1}) \ \in \ I\!R^{\nu-\nu_\alpha} \\
& v = (v_{\nu_\alpha,1}, \ldots, v_{\nu-1,\nu}) \ \in \ I\!R^{(\nu-\nu_\alpha)\nu} \\
& w \ \in \ I\!R.
\end{aligned}
$$

This equivalent reformulation of $DSqR^{\nu}$ involves $m + (\nu - \nu_\alpha)(\nu + 1) + 1$ variables and $2(\nu - \nu_\alpha)\nu + \nu$ inequality constraints. In practice, with the probability level α being set close to 1, $\nu_\alpha = \lceil \nu\alpha \rceil$ may be close to the number of observations ν. Consequently, the linear programming problem P^{ν}_{LP} becomes large-scaled when the sample size ν is large and decomposition algorithms may be needed.

Alternatively, we consider next a numerical integration-based scheme that avoids some auxiliary variables and constraints, and also handles the situation where the distribution of (X^{ν}, Y^{ν}) is not uniformly discrete.

2. Numerical Integration

The integral in $DSqR^{\nu}$ is easily approximated using standard numerical integration techniques. Suppose that the interval $[\alpha, 1]$ is divided into μ subintervals, where $\alpha \leq \beta_0 < \beta_1 < \ldots < \beta_{\mu-1} < \beta_{\mu} \leq 1$ and $w_i \geq 0, i = 0, 1, ..., \mu$, are factors specific to the integration scheme. An approximation of $DSqR^{\nu}$ then takes the form

Problem:

$$P^{\nu,\mu,w}_{\text{Num}}: \qquad \min_{c\in\boldsymbol{R}^m} \quad \frac{1}{1-\alpha}\sum_{i=0}^{\mu} w_i \bar{q}_{\beta_i}(Z_0^\nu(c)) \;-\; E\left[Z_0^\nu(c)\right].$$

For large μ, an optimal solution of problem $P^{\nu,\mu,w}_{\text{Num}}$ is close to that of $DSqR^\nu$, as seen next, under conditions that are satisfied by essentially all commonly used numerical integration schemes.

Proposition III.1. *Suppose that for any continuous function $g : [\alpha, 1] \to I\!R$, a numerical integration scheme with discretization points $\alpha \leq \beta_0 < \beta_1 < \ldots < \beta_{\mu-1} < \beta_\mu \leq 1$ and weights $w_i \geq 0, i = 0, 1, ..., \mu$, satisfies*

$$\left|\sum_{i=0}^{\mu} w_i g(\beta_i) - \int_\alpha^1 g(\beta)d\beta\right| \to 0$$

as $\mu \to \infty$. Let $\{\bar{c}^{\nu,\mu}\}_{\mu=1}^\infty$ be a sequence of optimal solutions of $P^{\nu,\mu,w}_{\text{Num}}$ under this numerical integration scheme. Then, every accumulation point of $\{\bar{c}^{\nu,\mu}\}_{\mu=1}^\infty$ is an optimal solution of $DSqR^\nu$.

Proof:

For any $c \in I\!R^m$, $\bar{q}_\beta(Z_0^\nu(c))$ is finite and continuous as a function of β. Consequently, the assumption on the numerical integration scheme applies and the objective function of $P^{\nu,\mu,w}_{\text{Num}}$ converges pointwise to that of $DSqR^\nu$, as $\mu \to \infty$.

The objective functions are also finite and convex in c, which follows directly from the convexity of $\bar{q}_\alpha$ on $\mathcal{L}^2$ and the affine form of Z_0^ν as a function of c. Consequently, by Theorem 7.17 in Rockafellar and Wets (1998), the objective function of $P^{\nu,\mu,w}_{\text{Num}}$ epiconverges to that of $DSqR^\nu$ and the conclusion follows from Theorem 7.31 in Rockafellar and Wets (1998).

□

While specialized solvers, such as Portfolio Safeguard in American Optimal Decisions, Inc. (2011), handle $P^{\nu,\mu,w}_{\text{Num}}$ directly with little difficulty under many circumstances, the problem is typically nonsmooth and standard nonlinear programming

solvers may fail. However, following a simple reformulation of $P^{\nu,\mu,w}_{\text{Num}}$, utilizing equation (II.7), we obtain the equivalent linear program formally stated below, where we assume for convenience that $\beta_\mu < 1$.

Problem:

$P^{\nu,\mu,w}_{\text{Num,LP}}$:

$$\begin{aligned}
\min_{c,u,v} \frac{1}{1-\alpha}\sum_{i=0}^{\mu} w_i \left(u_i + \frac{1}{1-\beta_i}\sum_{j=1}^{\nu} p^j v_{ij}\right) \quad - \quad & \sum_{j=1}^{\nu} p^j (y^j - \langle c, h(x^j)\rangle) \\
\text{s.t.} \qquad y^j - \langle c, h(x^j)\rangle - u_i \quad &\leq \quad v_{ij}, \quad i = 0, 1, ..., \mu \ , j = 1, ..., \nu \\
0 \quad &\leq \quad vij, \quad i = 0, 1, ..., \mu \ , j = 1, ..., \nu \\
c \quad &\in \quad I\!R^m \\
u = (u_0, u_1, ..., u_\mu) \quad &\in \quad I\!R^{\mu+1} \\
v = (v_{0,1}, ..., v_{\mu,\nu}) \quad &\in \quad I\!R^{(\mu+1)\nu}
\end{aligned}$$

If $\beta_\mu = 1$, then a straightforward modification is required based on the fact that $\bar{q}_1(Z^\nu_0(c)) = \max_{j=1,2,...,\nu} y^j - \langle c, x^j\rangle$. This linear program consists of $m+\mu+1+\nu(\mu+1)$ variables and $2\nu(\mu+1)$ constraints, which may be substantially less than what follows from the analytical integration approach for large ν. Here we assume that the weights $w_i \geq 0, i = 0, 1, ..., \mu$, are given and therefore not accounted for in the complexity analysis results. For example, in Chapter IV we assume that the $\mu + 1$ subintervals have the same weights. And in practice, we find that a moderately large μ suffices as shown in the numerical examples discussed in the same chapter.

B. DUAL METHODS

We now turn to a distinct perspective towards the alternative superquantile regression problem *DSqR*. We use the theory of the dualization of risk to build a

dual problem as described in the next subsection. We then solve this new problem using different algorithms, as seen in Subsections III.B.2 through III.B.4.

1. Dualization of Risk

We start this subsection by recalling the risk measure $\bar{\mathcal{R}}_\alpha$ corresponding to the superquantile as the statistic. According to equation (II.18), the measure of deviation for our superquantile-based quadrangle is described as follows

$$\bar{\mathcal{D}}_\alpha(Y) = \bar{\mathcal{R}}_\alpha(Y) - E\left[Y\right] = \frac{1}{1-\alpha}\int_\alpha^1 \bar{q}_\beta(Y)d\beta - E\left[Y\right] = \bar{\bar{q}}_\alpha(Y) - E\left[Y\right], \quad \text{(III.3)}$$

where $\bar{\mathcal{R}}_\alpha(Y) = \bar{\bar{q}}_\alpha(Y)$ is the risk measure for which we build the dual.

Next we turn to the dualization of risk measures and derive results that we can apply to our deviation-based superquantile regression problem $DSqR$. By the Envelope Theorem in Rockafellar and Uryasev (2013), an alternative formula for a positively homogeneous regular risk measure $\mathcal{R}(\cdot)$ is given by its dual representation, described as follows

$$\mathcal{R}(Y) = \sup_{Q\in\mathcal{Q}}\{E\left[YQ\right]\}, \quad \text{(III.4)}$$

where $\mathcal{Q}$ is a nonempty closed convex set that is to the risk envelope associated with $\mathcal{R}$. For $Y \in \mathcal{L}^2$ and $\alpha \in (0,1)$, a Q^Y that maximizes

$$\sup_{Q\in\mathcal{Q}}\{E\left[YQ\right]\}$$

is called a risk identifier. If Q^Y is a risk identifier, then obviously

$$\mathcal{R}(Y) = E[YQ^Y]. \quad \text{(III.5)}$$

Clearly, when we have a risk measure $\mathcal{R}(Y) = E[Y]$, we get $Q \equiv 1$. And for $\mathcal{R}(Y) = \sup Y$, we obtain $Q \in \{Q \in \mathcal{L}^2 \mid Q \geq 0, E[Q] = 1\}$.

For the general treatment of risk identifiers, we refer to Rockafellar and Royset (2014c). We consider the case where Ω is finite and $P(\{\omega\}) > 0$, for $\omega \in \Omega$, to avoid technical issues regarding measurability. We let $\Omega_\beta(Y) = \{\omega \in \Omega \mid Y(\omega) = q_\beta(Y)\}$, for

$\beta \in (0,1)$, and $F_Y^-(y)$ denote the left-continuous point of the cumulative distribution function F_Y.

Below we derive a risk identifier formula for the superquantile at Y and probability level $\beta \in (0,1)$.

Proposition III.2. *(Rockafellar & Royset, 2014c) For $\beta \in (0,1)$ and $Y \in \mathcal{L}^2$, a risk identifier for $\bar{q}_\beta(Y)$ is given by*

$$Q_\beta^Y(\omega) = \begin{cases} \frac{1}{1-\beta} & \text{if } Y(\omega) > q_\beta(Y), \\ r_\beta(\omega) & \text{if } Y(\omega) = q_\beta(Y), \\ 0 & \text{otherwise,} \end{cases}$$

with $0 \leq r_\beta(\omega) \leq 1/(1-\beta)$ for $\omega \in \Omega$ such that

$$\int_{\Omega_\beta(Y)} r_\beta(\omega) dP(\omega) = \frac{F_Y(q_\beta(Y)) - \beta}{1-\beta}. \tag{III.6}$$

We now turn to the risk identifier for our choice of measure of risk in problem $DSqR$, the α-second-order superquantile. We interpret 0 times $-\infty$ as zero. Let $\bar{Q}_\alpha^Y$ be a risk identifier for $\bar{\bar{q}}_\alpha(Y)$.

Proposition III.3. *(Rockafellar & Royset, 2014c) Suppose that Y has a discrete distribution with ν possible realizations. Then, for $\alpha \in (0,1)$ and $Y \in \mathcal{L}^2$, a risk identifier of $\bar{\bar{q}}_\alpha(Y)$, is given by*

$$\bar{Q}_{\alpha}^{Y}(\omega) = \begin{cases} \frac{1}{1-\alpha}\log\frac{1-\alpha}{1-F_Y(Y(\omega))} & \text{if } \alpha < F_Y^-(Y(\omega)) = F_Y(Y(\omega)) < 1 \\ \frac{1}{1-\alpha}\left[\log\frac{1-\alpha}{1-F_Y^-(Y(\omega))}+1\right] & \text{if } \alpha < F_Y^-(Y(\omega)) < F_Y(Y(\omega)) = 1 \\ \frac{1}{1-\alpha}\Big[\log\frac{1-\alpha}{1-F_Y^-(Y(\omega))}+1 & \\ \quad + \frac{1-F_Y(Y(\omega))}{F_Y(Y(\omega))-F_Y^-(Y(\omega))}\log\frac{1-F_Y(Y(\omega))}{1-F_Y^-(Y(\omega))}\,\Big] & \text{if } \alpha < F_Y^-(Y(\omega)) < F_Y(Y(\omega)) \\ \frac{1}{1-\alpha}\left[\frac{F_Y(Y(\omega))-\alpha}{F_Y(Y(\omega))-F_Y^-(Y(\omega))}\right] & \text{if } F_Y^-(Y(\omega)) < \alpha < F_Y(Y(\omega)) = 1 \\ \frac{1}{1-\alpha}\Big[\frac{F_Y(Y(\omega))-\alpha}{F_Y(Y(\omega))-F_Y^-(Y(\omega))} & \\ \quad + \frac{1-F_Y(Y(\omega))}{F_Y(Y(\omega))-F_Y^-(Y(\omega))}\log\frac{1-F_Y(Y(\omega))}{1-\alpha}\,\Big] & \text{if } F_Y^-(Y(\omega)) \le \alpha \le F_Y(Y(\omega)) \\ & \quad \text{and } F_Y^-(Y(\omega)) < F_Y(Y(\omega)) \\ 0 & \text{otherwise.} \end{cases}$$

In view of Theorem 4.13 in Rockafellar and Royset (2014c), and equations (III.3) and (III.4), we are now able to build a dual method to solve the Deviation-based Superquantile Regression Problem $DSqR$.

Consider the risk identifier $\bar{Q}_{\alpha}^{Z_0(c)}$ of $\bar{\bar{q}}_{\alpha}(Z_0(c))$, as defined in Proposition III.3, for a probability level $\alpha \in (0,1)$. Then, according to equation (III.3), we have that

$$\begin{aligned} \bar{\mathcal{D}}_{\alpha}(Z_0(c)) &= \bar{\mathcal{R}}_{\alpha}(Z_0(c)) - E\left[Z_0(c)\right] \\ &= \bar{\bar{q}}_{\alpha}(Z_0(c)) - E\left[Z_0(c)\right] \\ &= E[Z_0(c)\bar{Q}_{\alpha}^{Z_0(c)}] - E\left[Z_0(c)\right]. \end{aligned} \tag{III.7}$$

And we are able to define the objective function of this new problem as follows

$$\begin{aligned} f(c) &= \frac{1}{\nu}\sum_{i=1}^{\nu} Z_0(c)^{(i)}\bar{Q}_{\alpha}^{Z_0(c)}(i) - \frac{1}{\nu}\sum_{j=1}^{\nu} Z_0(c)^{j} \\ &= \frac{1}{\nu}\sum_{i=1}^{\nu}\left(y^{(i)} - \langle c, h(x^{(i)})\rangle\right)\bar{Q}_{\alpha}^{Z_0(c)}(i) - \frac{1}{\nu}\sum_{j=1}^{\nu}\left(y^{j} - \langle c, h(x^{j})\rangle\right). \end{aligned} \tag{III.8}$$

where $Z_0(c)^{(i)}$ is the i^{th}-ordered value of $Z_0(c)$. The evaluation of the objective function requires the computation of $\bar{Q}_{\alpha}^{Z_0(c)}$. According to Proposition III.3, this implies

sorting vector $Z_0(c)$ for a given c to obtain its cumulative distribution function and only then evaluate $\bar{Q}_\alpha^{Z_0(c)}$, using the same sorting as for $Z_0(c)^{(i)}$. A subgradient of $f(c)$ is then easily computed as follows

$$\nabla f(c) = -\frac{1}{\nu}\sum_{i=1}^{\nu} h\left(x^{(i)}\right)\bar{Q}_\alpha^{Z_0(c)}(i) + \frac{1}{\nu}\sum_{j=1}^{\nu} h\left(x^{j}\right), \tag{III.9}$$

with $h\left(x^{(i)}\right)$ maintaining the same ordering as in $Z_0(c)^{(i)}$ used in (III.8).

The Approximate Dualization-of-risk Superquantile Regression Problem D^ν is defined as:

Problem:

$$D^\nu: \qquad \min_{c\in R^m} \quad \bar{\mathcal{D}}_\alpha\left(Z_0^\nu(c)\right) = \frac{1}{\nu}\sum_{i=1}^{\nu} Z_0^\nu(c)^{(i)}\bar{Q}_\alpha^{Z_0^\nu(c)}(i) - \frac{1}{\nu}\sum_{j=1}^{\nu} Z_0^\nu(c)^{j},$$

with $\bar{c}_0^\nu$ being obtained by setting $\bar{c}_0^\nu = \bar{q}_\alpha(Z_0^\nu(\bar{c}^\nu))$, and $\bar{Q}_\alpha^{Z_0^\nu(c)}$ given by Proposition III.3.

We now turn to the implementation of these results. In the next subsections we present three algorithms that are well known. First we start with a simple algorithm, the subgradient method, and then move to an heuristic algorithm, the coordinate descent method, and finish off with the cutting plane method. There are obviously many other possible algorithms we could implement when solving the dual methods, but we omit such investigation and only discuss these three as examples.

2. Subgradient Method

The subgradient method was originally developed by Naum Z. Shor and others in the 1960s and 1970s; see Shor (1985). It is a simple algorithm that can be implemented for solving a wide variety of problems, such as the minimization of nondifferentiable convex functions.

The subgradient method is an iterative algorithm that aims to minimize a convex function f, by iteratively obtaining a new c^{k+1} according to the following

scheme

$$c^{k+1} = c^k - \delta^k \nabla f(c^k), \tag{III.10}$$

where $\nabla f(c^k)$ is any subgradient of f evaluated at c^k, and δ^k is the stepsize used in iteration k. As a downside, this algorithm is not a descent method and it is possible to obtain increased objective function values in any iteration, therefore we need to store the best obtained objective function value by setting $f^k_{best} = \min\left\{f^{k-1}_{best}, f^k_{best}\right\}$. In fact, if we obtain the best function value so far in iteration k, we also need to store $i^k_{best} = k$. This way we guarantee to have $f^k_{best} = \min\left\{f(c^1), f(c^2), \ldots, f(c^k)\right\}$ stored for later use.

There are obviously many rules to define the stepsize used in algorithm **SM**, as we describe below. For example, one could use a step with constant length instead, $\delta_k = \delta/\|\nabla f(c^k)\|_2$, so that $\|c^{k+1} - c^k\|_2 = \delta$, or perhaps a diminishing stepsize, such as $\delta_k = \gamma_1/(k + \gamma_2)$, with γ_1 and γ_2 being some positive scalars. The importance of the right choice of stepsize δ^k becomes more aparent when we discuss computational performances, later in Section III.C.

We now formally describe the subgradient method.

Algorithm SM:

Step 0. Choose an initial guess $c^0 \in I\!R^m$. Set $k := 0$.
Initialize $f^0_{best} := \infty$, and $i^0_{best} := 0$.

Step 1. Compute $f(c^k)$ and $\nabla f(c^k)$, using Equations (III.8) and (III.9), respectively.
If $\nabla f(c^k) = 0$, then c^k is an optimal solution, stop.

Step 2. Set $f^k_{best} = \min\left\{f^{k-1}_{best}, f(c^k)\right\}$, and let $i^k_{best} = k$ if $f(c^k) = f^k_{best}$.

Step 3. Choose stepsize δ^k, with $\delta^k > 0$.

Step 4. Define $c^{k+1} = c^k - \delta^k \nabla f(c^k)$.
Replace k by $k + 1$ and go to Step 1.

3. Coordinate Descent Method

The coordinate descent method is an heuristic algorithm that is simple to implement. In this method, the objective function is minimized along one coordinate direction per iteration and a cycle is complete when all coordinates have been utilized in this process. Although we could define any permutation of coordinates as the order for the coordinate search, we will use the cyclical order for simplification. We benefit from the possibility of computing the subgradient of the objective function, as defined in (III.9), to perform line search in each coordinate direction.

We now formally describe the coordinate descent method.

Algorithm CDM:

Step 0. Start with an initial guess $c^0 \in I\!R^m$.

Set the cycle counter $k := 1$.

Step 1. Choose coordinate 1 and compute $c_1^k \in \text{argmin}_{c_1} f(c_1, c_2^{k-1}, \ldots, c_m^{k-1})$.

Step 2. Choose coordinate 2 and compute $c_2^k \in \text{argmin}_{c_2} f(c_1^k, c_2, \ldots, c_m^{k-1})$.

...

Step m. Choose coordinate m and compute $c_m^k \in \text{argmin}_{c_m} f(c_1^k, c_2^k, \ldots, c_m^{k-1})$.

Replace cycle k by $k+1$ and go to Step 1.

This algorithm terminates according to the threshold tolerance $\epsilon > 0$, inputted by the decision maker. For simplicity, we use the formula $f(c^{k-1}) - f(c^k) \leq \epsilon$ as our stopping criteria.

4. Cutting Plane Method

We finish Section III.B by describing the third algorithm we implement in the numerical examples, in Chapter IV: the cutting plane method, which is guaranteed to achieve an optimal solution if one exists.

The idea behind this algorithm is to solve an approximate linear program each iteration. The cutting plane method starts off with our original unconstrained

problem and with every iteration we obtain a cut to the feasible region that we add as a new constraint for the following linear program. So it approximates the feasible region by a finite set of closed half spaces and solves a sequence of approximating linear programs until the optimal solution is found. As we notice, the size of the linear program grows with the number of iterations and becomes rather slow for a larger number of variables.

The cutting plane method is usually used in integer or mixed integer linear programming problems but is also very popular when applied to convex minimization problems whenever the objective function value and its subgradient are easily computed, as we describe in detail below. Consider our minimization problem

$$\min_{c\in R^m} \quad f(c) = \frac{1}{\nu}\sum_{i=1}^{\nu} Z_0(c)^{(i)}\bar{Q}_{\alpha}^{Z_0(c)}(i) - \frac{1}{\nu}\sum_{j=1}^{\nu} Z_0(c)^{j}.$$

Using $\nabla f(c^0)$, see Equation (III.9), at an initial guess $c^0 \in I\!R^m$, we are able to build a relaxation to our problem, as follows

$$\begin{array}{ll} \min\limits_{\xi,c} & \xi \\ \text{s.t.} & f(c^0) + \nabla f(c^0)^\top \left(c - c^0\right) \ \leq \ \xi, \end{array} \tag{III.11}$$

with $\xi \in I\!R$ being a dummy variable. If we keep adding a new constraint per iteration k, as in (III.11), but now applied to the obtained optimal solution $\bar{c}^{k-1}$, we construct the linear programming problem with K constraints, where K denotes the total number of iterations,

$$\begin{array}{ll} \min\limits_{\xi,c} & \xi \\ \text{s.t.} & f(c^k) + \nabla f(c^k)^\top \left(c - c^k\right) \ \leq \ \xi, \qquad k = 0, \ldots, K. \end{array}$$

We now formally state the cutting plane method.

Algorithm CPM:

Step 0. Start with an initial guess $c^0 \in I\!R^m$. Set $k := 0$.

Step 1. Compute $f(c^k)$ and $\nabla f(c^k)$, using Equations (III.8) and (III.9), respectively.

If $\nabla f(c^k) = 0$, then c^k is an optimal solution, stop.

Step 2. Solve the Linear Program

$$\begin{aligned} \min_{\xi,c} \quad & \xi \\ \text{s.t.} \quad & f(c^i) + \nabla f(c^i)^\top (c - c^i) \ \leq \ \xi, \quad i = 0, \ldots, k. \end{aligned}$$

Step 3. Get obtained optimal solution $\bar{c}$ from Step 2 and set $c^{k+1} = \bar{c}$.

Replace k by $k + 1$ and go to Step 1.

In the next section, we compare computational performances of the algorithms we present in CHapters III.A and III.B. We also compare these complexity results with least squares and quantile regression in order to understand how good these presented computational methods are.

C. COMPLEXITY

In the previous two sections we present different computational methods for the superquantile regression problem. When implementing these methods, it is useful to know how efficient and costly they are. In this section, we compare primal versus dual methods in terms of worst-case complexity, and analyze the computational performances of least squares and quantile regressions.

1. Least Squares Regression

In the case of least squares regression we have a system with ν linear equations, due to the ν observations in the data set, and $m + 1$ unknown coefficients, $(c_0, c_1, \ldots, c_m)$. We let X be a design matrix of dimension ν by $(m + 1)$, with all

elements in the first column being set equal to 1 in order for us to be able to include the intercept c_0 in the regression model.

Then the best fitting coefficients $(\bar{c}_0, \bar{c})$ are the ones obtained by solving the quadratic minimization problem

$$\min_{(c_0,c)\in R^{m+1}} \sum_{i=1}^{\nu} \left(y^i - c_0 - \sum_{j=1}^{m} X_{ij} c_j \right)^2,$$

and, in matrix notation, are equal to

$$(\bar{c}_0, \bar{c}) \;=\; (X^\top X)^{-1} X^\top y. \tag{III.12}$$

In terms of computational cost this algorithm implies: multiplying $X^\top$ by X, which takes $O(\nu(m+1)^2)$ arithmetic operations; multiplying $X^\top$ by y, which takes another $O(\nu(m+1))$ arithmetic operations; computing the LU factorization of $(X^\top X)$, which takes another $O((m+1)^3)$ arithmetic operations; and finally multiplying $(X^\top X)^{-1}$ by $(X^\top y)$, which takes $O((m+1)^2)$. So overall the running time of this procedure is $O(\nu m^2)$, assuming of course that $\nu > m$ and X is a full rank matrix.

2. Quantile Regression

As discussed in Subsection II.A.3, the quantile regression function is obtained by minimizing the (scaled) Koenker-Bassett error measure (Koenker, 2005). This problem can be rewritten as a linear program as follows

$$\begin{aligned}
\min_{c_0,c,u,v} \quad & \alpha\; 1_\nu^\top\; u + (1-\alpha)\; 1_\nu^\top\; v \\
\text{s.t.} \quad & c_0 + \langle c, h(x^i)\rangle + u_i - v_i \;=\; y^i, \quad i = 1, \ldots, \nu \\
& c_0 \;\in\; I\!R \\
& c \;\in\; I\!R^m \\
& u = (u_1, \ldots, u_\nu) \;\in\; I\!R^\nu \\
& v = (v_1, \ldots, v_\nu) \;\in\; I\!R^\nu,
\end{aligned}$$

where $1_\nu^\top$ denotes a transposed ν-dimensional vector of ones. This linear program has a total of $2\nu + m + 1$ number of variables and ν number of equality constraints. For

us to be able to proceed with the computational performance analysis, we need to transform the problem into standard form. Summarizing we then have $2(2\nu + m + 1)$ variables and ν equality constraints.

Solving this linear program by means of an interior point method takes $O((4\nu + 2m + 2)^{3.5})$ operations to produce a solution. The path following algorithm is one of such interior point methods. Monteiro and Adler (1989) refined the path following algorithm to converge in $O(\sqrt{2(2\nu + m + 1)}\log(\epsilon_0/\epsilon))$ iterations by reducing the duality gap from ϵ_0 to ϵ, with $O\left((4\nu + 2m + 2)^3\right)$ arithmetic operations per iteration.

The quantile regression implementation takes a total of $O\left(\nu^{3.5}\log(\epsilon_0/\epsilon)\right)$, assuming that ν is larger than m. Specialized algorithms (see for example Koenker, 2005) improve on this solution approach, but further discussions are beyond the scope of this dissertation.

3. Superquantile Regression – Primal Methods

Let us start with the analytical integration presented in Subsection III.A.1. We determine the computational performance of this method when the resulting linear program is solved using an interior point method.

In order to determine the computational performance of problem P^{ν}_{LP}, we need to transform P^{ν}_{LP} into a standard form linear programming problem. After this transformation, we have $2\left[(\nu - \nu_\alpha)(\nu + 1) + m + 1\right] + \nu$ variables and $(\nu - \nu_\alpha)\nu + \nu$ equality constraints. Since $\nu_\alpha = \lceil \nu\alpha \rceil$, with α being usually close to 1, ν_α is almost as big as the number of observations ν in the data set.

As done with the computational performance in the quantile regression case, we use the convergence results we find in Monteiro and Adler (1989). The primal method using analytical integration takes a total of $O\left(\nu^7\log(\epsilon_0/\epsilon)\right)$.

Let us now turn to the numerical integration method described in Subsection III.A.2. Problem $P^{\nu,\mu,w}_{\text{Num}}$ is a linear program with $m + \mu + 1 + \nu(\mu + 1)$ variables and $2\nu(\mu+1)$ inequality constraints. After transforming $P^{\nu,\mu,w}_{\text{Num}}$ into a standard form linear program, we have $2(\mu\nu + \mu + \nu + m + 1)$ variables and $(\mu + 1)\nu$ equality constraints

in the primal method with numerical integration. Since the number of observations ν and integration subintervals μ are both usually large numbers and we disregard the inputted weights in our complexity analysis, the implementation of the primal methods for superquantile regression takes a total of $O\left(\mu^{3.5}\nu^{3.5}\log(\epsilon_0/\epsilon)\right)$.

4. Superquantile Regression – Dual Methods

We now compare the computational performance of the dual methods. Since in the numerical examples we implement the subgradient method using a constant stepsize rule, we analyze the computational performance of this algorithm under this circumstance.

Let $d(c^0) = \min_{\bar{c}\in R^m} \|c^0 - \bar{c}\|$ be the distance between the initial guess c^0 and the optimal solution $\bar{c}$. And let $\{c^k\}$ be the sequence generated by the subgradient method, with the stepsize δ^k fixed at some positive constant δ, with $k \in \{1, 2, \ldots, K\}$.

Then, according to Proposition 6.3.3 in Bertsekas (2009), for any scalar $\epsilon > 0$, we have that

$$\min_{0\leq k\leq K} f(c^k) \leq f(\bar{c}) + \frac{\delta u^2 + \epsilon}{2}, \tag{III.13}$$

where K is given by

$$K = \left\lfloor \frac{d(c^0)^2}{\delta\epsilon} \right\rfloor, \tag{III.14}$$

with u being the upper bound on the norm of $\nabla f(c^k) \in \partial f(c^k), \forall k \geq 0$. The number of iterations is independent of the number of variables in the problem. The most costly operation of this algorithm in our case is the computation of $\nabla f(c^k)$ at any given iteration k. Since the vector $Z_0(c^k)$ needs to be sorted in order to compute $\nabla f(c^k)$, as stated in equation (III.8), the subgradient method takes $O(\nu \log \nu)$ operations per iteration. Note that by establishing $\delta = \epsilon/u$, we can obtain an ϵ-optimal solution in $O(1/\epsilon^2)$ iterations. So the subgradient method takes a total of $O((1/\epsilon^2)\nu \log \nu)$.

We note that we present the complexity result for the slowest of the described dual methods. Implementing the Nesterov's optimal method (see Nesterov, 1983) improves the obtained result for a total of $O((1/\epsilon)\nu \log \nu)$.

These results show that dual methods are not much slower than solving for least squares regression and such a conclusion is promising for superquantile regression.

In the next chapter we present a series of numerical examples that allow us to compare runtimes of the various algorithms.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. NUMERICAL EXAMPLES

In this chapter, we illustrate superquantile regression in several numerical examples. We start with a simple example that allows us to compare computational methods in terms of runtimes, solution vectors and function values. Then we apply superquantile regression to the well-known data sets, Engel data and Brownlee stack loss plant data, and compare the obtained superquantile regression models to least squares and quantile regression functions. In the fourth example, we apply superquantile regression to an investment analysis problem taken from a case study of the Portfolio Safeguard documentation (American Optimal Decisions, Inc., 2011). The fifth and sixth examples address military applications, the first concerning U.S. Navy helicopter pilots and the second Portuguese Navy submariners, and in both examples their mission employment. We then show an example that arises in uncertainty quantification of a rectangular cross section of a short structural column under uncertain yield stress and uncertain loads. Finally we revisit the first example in order to address the issue of superquantile tracking. We experiment different regression models. We compare the obtained solution vectors, coefficients of determination and adjusted coefficients of determinations, and implement Cook's distances applied to superquantile regression.

Computations are mostly carried out in Matlab version 7.14 on a 2.26 GHz laptop with 8.0 GB of RAM using Windows 7. However we implement both least squares and quantile regression in R programming language (R Development Core Team, 2008). Specifically for solving the superquantile regression problem with a numerical integration scheme $P_{\text{Num}}^{\nu,\mu,w}$, we use Portfolio Safeguard in Matlab, by American Optimal Decisions, Inc. (2011), with VAN as the optimization solver. Since we assume the subintervals are equally likely when solving for the primal method using numerical integration schemes, from now on we denote problem $P_{\text{Num}}^{\nu,\mu,w}$ by $P_{\text{Num}}^{\nu,\mu}$ instead, and assume $w_i = 1/\mu$, for $i = 0, 1, ..., \mu$. When solving the primal method with

analytical integration, P^{ν}_{LP}, and the dual method, D^{ν}, we employ GAMS version 23.7 with the CPLEX 12.3 solver.

A. COMPUTATIONAL COST

We start by considering a loss random variable

$$Y = X_1 + X_2\epsilon,$$

where ϵ is a standard normal random variable and $X = (X_1, X_2)$ is uniformly distributed on $[-1, 1] \times [0, 1]$, with ϵ, X_1, and X_2 independent. We consider a regression function of the form $f(x) = c_0 + c_1x_1 + c_2x_2$ and set $\alpha = 0.90$. This simple example serves as a comparison tool between computational methods and their performances, as well as the obtained approximate solution vectors, since we know the conditional superquantile, which in this case is $(c_0, c_1, c_2) = (0, 1, 1.755)$. The detailed explanation can be seen in Section IV.H. We give an initial guess $(c^0_1, c^0_2) = (0, 0)$ when implementing the dual methods, and $\bar{c}_0$ is consecutively computed utilizing the regression vector $(\bar{c}_1, \bar{c}_2)$ obtained by the implemented algorithm.

We first examine the computational effort required to obtain an approximate regression vector. Table 1 shows computing times for solving problem P^{ν}_{LP} for increasingly larger sample sizes ν obtained by independent draws from (ϵ, X_1, X_2). While the results correspond to single instances of P^{ν}_{LP}, the times vary little between two instances of the same size and the computing times are therefore representative. As

ν	100	200	300	400	500	600	700	800	900	1000	1500	2000
Time	0	0	2	6	17	32	45	65	163	174	996	2,972

Table 1. Example A: Computing times (sec.) for solving P^{ν}_{LP} for increasingly larger sample sizes ν.

predicted from the theoretical results discussed at the end of Subsection III.A.1, the computing time grows rapidly as the sample size ν increases. In addition to the

inconvenience of long computing times, memory requirements become problematic. Therefore solving P^{ν}_{LP} for large sample sizes is challenging, if not impossible, and we examine alternative approaches.

Second, we consider the alternative primal method approach based on solving $P^{\nu,\mu}_{\text{Num}}$. While this approach introduces a numerical integration error, Proposition III.1 ensures that the error is negligible for large μ. In fact, as we see next empirically, moderately large μ suffices for probability levels α close to one. Moreover, the substantial reduction in problem size, as compared to that of P^{ν}_{LP}, reduces computing times dramatically.

Since $\bar{q}_{\beta}(Z^{\nu}_{0}(c))$ may be nonsmooth as a function of β, standard numerical integration error bounds may not apply. However, since $\bar{q}_{\beta}(Z^{\nu}_{0}(c))$ is continuous and nondecreasing as a function of β, the use of left-endpoint and right-endpoint numerical integration rules in $P^{\nu,\mu}_{\text{Num}}$ provide lower and upper bounds on the optimal value of $DSqR^{\nu}$, respectively. Table 2 shows solution vectors (c_0, c_1, c_2) for $\mu = 100$, $\mu = 1000$, and $\mu = 5000$ integration subintervals, when we implement left-endpoint, right-endpoint, and Simpson's numerical integration schemes, for sample sizes of $\nu = 100$, $\nu = 1000$, $\nu = 10000$, and $\nu = 100000$.

For $\nu = 100$, we notice that the solutions are insensitive to the numerical integration rule as well as the subintervals μ specific to the integration scheme. The obtained solutions are essentially identical to the regression vector obtained from P^{ν}_{LP}; see Row 2 of Table 2. Here the superquantile coefficient of determination is $\bar{R}^2_{0.90} = 0.5683$ for all the presented cases, including P^{ν}_{LP}, which also supports the fact that the numerical integration rule does not affect the obtained solution. Each one of the solutions of $P^{\nu,\mu}_{\text{Num}}$ for $\nu = 100$ is obtained quickly, in about 0.08 to 8 seconds for $\mu = 100$, $\mu = 1000$, and $\mu = 5000$; see the last column of Table 2. In this case, we clearly notice that $\mu = 100$ suffices and takes less than a tenth of a second to run.

When we increase the sample size ν, we start to notice that the solution vectors are slightly different but the magnitudes of these differences are small for subintervals

Problem	Integration Rule	ν	μ	c_0	c_1	c_2	Time
P_{LP}^{ν}	NA	100	NA	0.0630	1.0951	1.5841	0.05
$P_{\text{Num}}^{\nu,\mu}$	Left Endpoint	100	100	0.0630	1.0951	1.5841	0.07
	Right Endpoint	100	100	0.0630	1.0951	1.5841	0.08
	Simpson's	100	100	0.0630	1.0951	1.5841	0.09
	Left Endpoint	100	1000	0.0630	1.0951	1.5841	0.79
	Right Endpoint	100	1000	0.0630	1.0951	1.5841	0.83
	Simpson's	100	1000	0.0630	1.0951	1.5841	0.77
	Left Endpoint	100	5000	0.0630	1.0951	1.5841	7.81
	Right Endpoint	100	5000	0.0630	1.0951	1.5841	7.24
	Simpson's	100	5000	0.0630	1.0951	1.5841	7.10
P_{LP}^{ν}	NA	1000	NA	0.0680	1.0108	1.7322	174.24
$P_{\text{Num}}^{\nu,\mu}$	Left Endpoint	1000	100	0.0689	1.0112	1.7290	0.12
	Right Endpoint	1000	100	0.0658	1.0099	1.7398	0.13
	Simpson's	1000	100	0.0680	1.0108	1.7322	0.13
	Left Endpoint	1000	1000	0.0683	1.0112	1.7310	1.29
	Right Endpoint	1000	1000	0.0678	1.0106	1.7327	1.26
	Simpson's	1000	1000	0.0680	1.0108	1.7322	1.21
	Left Endpoint	1000	5000	0.0680	1.0109	1.7321	10.91
	Right Endpoint	1000	5000	0.0680	1.0108	1.7322	11.44
	Simpson's	1000	5000	0.0680	1.0108	1.7322	9.76
$P_{\text{Num}}^{\nu,\mu}$	Left Endpoint	10000	100	0.0835	1.0049	1.6374	0.58
	Right Endpoint	10000	100	0.0799	1.0050	1.6492	0.56
	Simpson's	10000	100	0.0818	1.0048	1.6429	0.56
	Left Endpoint	10000	1000	0.0820	1.0048	1.6423	5.91
	Right Endpoint	10000	1000	0.0816	1.0048	1.6435	5.00
	Simpson's	10000	1000	0.0818	1.0048	1.6430	5.27
	Left Endpoint	10000	5000	0.0818	1.0048	1.6428	28.93
	Right Endpoint	10000	5000	0.0817	1.0048	1.6431	32.72
	Simpson's	10000	5000	0.0818	1.0048	1.6429	29.12
$P_{\text{Num}}^{\nu,\mu}$	Left Endpoint	100000	100	0.8149	0.2484	1.1749	5.05
	Right Endpoint	100000	100	0.8242	0.2411	1.1572	4.55
	Simpson's	100000	100	0.8176	0.2454	1.1702	3.98
	Left Endpoint	100000	1000	0.8152	0.2462	1.1750	46.00
	Right Endpoint	100000	1000	0.8162	0.2454	1.1732	38.01
	Simpson's	100000	1000	0.8155	0.2459	1.1746	46.07
	Left Endpoint	100000	5000	0.8155	0.2460	1.1746	307.34
	Right Endpoint	100000	5000	0.8156	0.2458	1.1743	330.99
	Simpson's	100000	5000	0.8156	0.2459	1.1744	278.55

Table 2. Example A: Solution vectors and computing times (sec.) for varying number of observations ν, integration rules for solving $P_{\text{Num}}^{\nu,\mu}$ as well as number of integration subintervals μ.

of $\mu = 100$ and $\mu = 1000$. For numerical integration scheme implementations with $\mu = 5000$, these differences are almost inexistent, but the computing times are larger. Therefore the statistician should take this into consideration when selecting the number of subintervals for the numerical integration scheme. It is a tradeoff between obtaining better solutions versus computing times. Also we notice there is another issue we encounter when solving superquantile regression problems for sample sizes as large as 100000 observations. In Rows 31-39 of Table 2, we intentionally include the solution vectors for $\nu = 100000$ using the same number of subintervals as implemented in the other cases. The discrepancies in solution vectors are consequence of rounding errors and we refer to Borges (2011) for further details.

One detail that is not included in Table 2 is the coefficient of determination $\bar{R}^2_{0.90}$. For all the presented cases, the coefficient of determination takes the values of 0.4222, 0.3917, and 0.1029, for sample sizes $\nu = 1000$, $\nu = 10000$, and $\nu = 100000$, respectively. We notice that $\bar{R}^2_{0.90}$ decreases as we increase the size of the data sample, which means that the linear model $f(x) = c_0 + c_1 x_1 + c_2 x_2$ does not fully capture the variability of the data, as expected, and a study of other models may be warranted. However, we omit such an investigation.

As discussed in Chapter III, the dual method is another approach to solve the deviation-based superquantile regression problem which theoretically demonstrates potential for large sample sizes. Since numerical integration not only introduces a numerical integration error but also takes increasingly longer to run for increasing sample sizes, we proceed with the implementation of the dual methods.

Third, we solve the superquantile regression problem by implementing the dual methods of Section B in Chapter III, i.e., subgradient, coordinate descent, and cutting plane methods. Since defining the stepsize and tolerance for the three algorithms, as well as the maximum number of iterations in the specific case of the subgradient method, can be a difficult process, we establish the following input parameters for each algorithm as a natural choice for us to be able to compare all three methods. We

note that refining these parameters as well as implementing more efficient algorithms could return even better computing times but such an investigation is not the purpose here. Our goal with this example is to demonstrate the potential for dual methods. For the subgradient method, we fix the stepsize to a constant value, $\delta = 0.1$, and run the algorithm for 1000 iterations. In the case of the coordinate descent method, we include a tolerance of 10^{-12} and define 1000 as the maximum number of iterations. We implement the cutting plane method with a maximum of 1000 cuts, and a gap of 10^{-8}.

Table 3 shows the computing times needed for solving problem D^ν for increasingly large sample sizes ν implementing these algorithms. Here the computing times are also representative, for the same reason as in Table 1. As expected, the subgradi-

ν	Computing Times		
	Subgradient	Coordinate Descent	Cutting Plane
100	0.13	0.67	0.91
1000	0.34	1.20	2.07
5000	1.43	0.70	0.95
10000	2.88	0.88	3.00
25000	5.96	3.24	2.20
50000	11.92	8.30	1.73
75000	21.53	13.78	1.98
100000	27.38	19.20	1.78

Table 3. Example A: Computing times (sec.) for solving D^ν using different implementations of the dual methods for increasing sample sizes ν.

ent method is the slowest of the three algorithms for almost all the presented cases, especially for sample sizes greater than 1000 observations. In all the described dual methods and empirically, the computing times grow linearly with the sample size ν, with the cutting plane method having the smallest slope of the three, as shown in Figure 6.

In Figure 7, we picture the computing times for primal versus dual methods, in logarithmic scale. Here we choose to present the Simpson's integration rule with

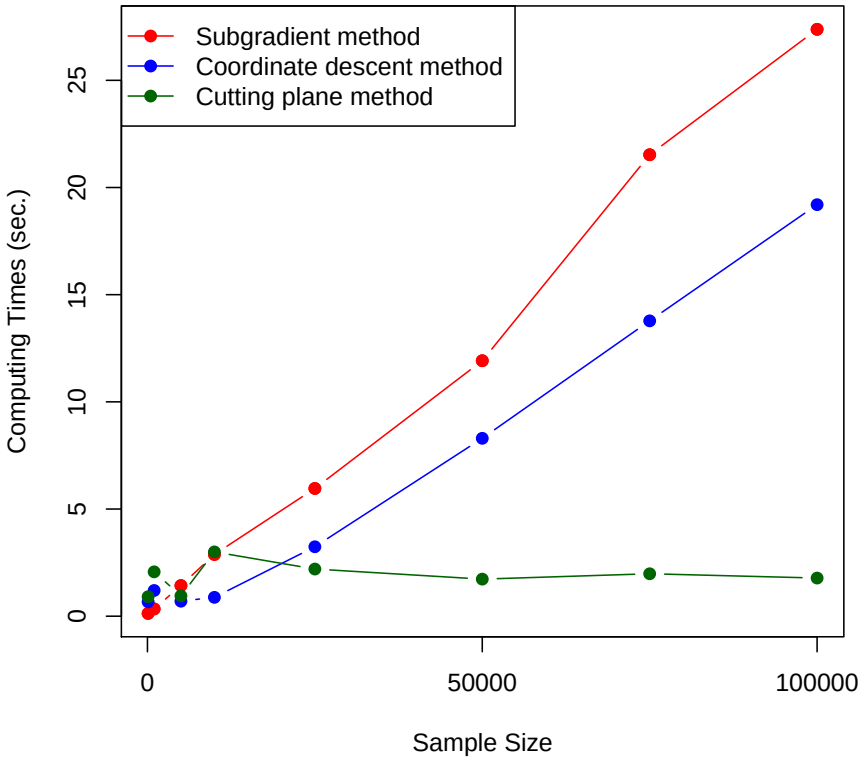


Figure 6. Example A: Computing times for solving D^{ν} with three different algorithms (subgradient, coordinate descent, and cutting plane methods), for increasing sample sizes ν.

$\mu = 1000$ subintervals and the dual methods algorithms with the input parameters as stated before. We clearly notice that implementing the cutting plane method improves the computational performance, especially for large sample sizes. Also, for larger samples sizes and smaller probability levels α, we certainly need to increase the number of integration subintervals μ.

We also compare the obtained solution vectors and corresponding objective function values; see Table 4. Again we note that it is not possible to solve P_{LP}^{ν} for sample sizes larger than $\nu = 1000$. We use Simpson's rule with $\mu = 1000$ intervals as the numerical integration scheme for all sample sizes. We realize that the obtained solution vectors are nearly identical.

Finally we analyze how changing the probability level α and the number of

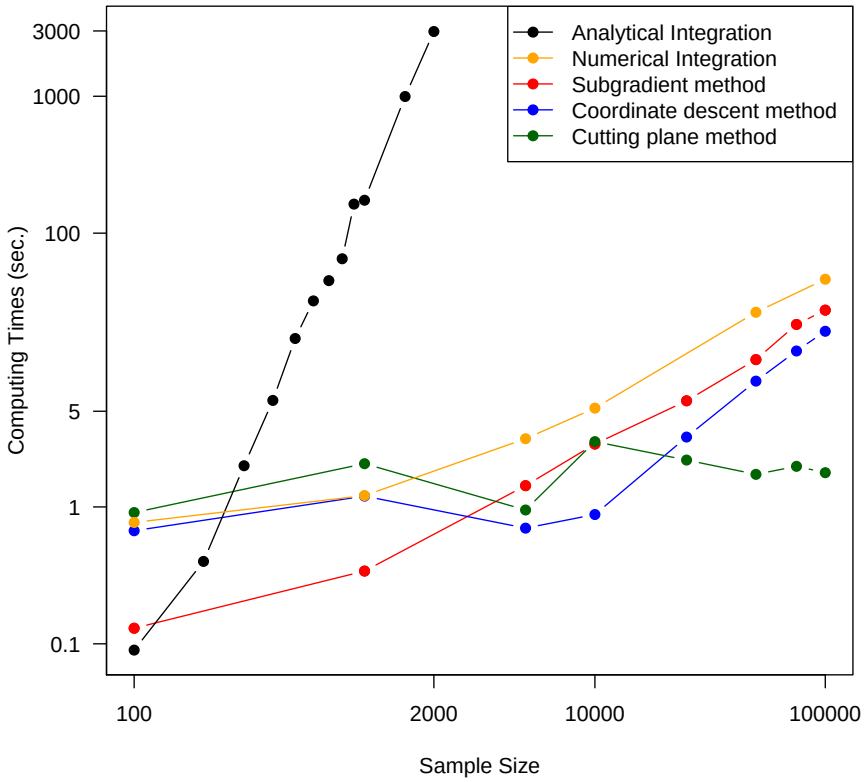


Figure 7. Example A: Primal versus dual methods computing times for increasing sample sizes ν, in logarithmic scale.

observations ν affect the computational performance of the implemented algorithms. Obviously, the primal methods are affected by changes in the probability level α since the integral in problem $DSqR^{\nu}$ is defined between α and 1. The smaller the value of α, the smaller $\nu_{\alpha} = \lceil \nu\alpha \rceil$ gets, and consequently the number of variables and inequality constraints in problem P^{ν}_{LP} increases due to the increased difference $(\nu - \nu_{\alpha})$. In the numerical integration schemes, the smaller α gets, the more subintervals μ are required to obtain same accuracy.

As we can observe in Table 5, the sample size ν influences the computing times of the subgradient method, but the probability level α does not produce such an effect. We note that the computing times for the sample sizes $\nu = 100$ and $\nu = 1000$ are not exactly the same for all the presented probability levels α. These values differ in the

Computational Method	ν	μ	c_0	c_1	c_2	Function Value	Time
Analytical Int.	100	NA	0.0630	1.0951	1.5841	0.844477	0.05
Numerical Int.	100	1000	0.0630	1.0951	1.5841	0.844477	0.77
Subgradient	100	NA	0.0673	1.1004	1.5699	0.844575	0.13
Coord. Descent	100	NA	0.0631	1.0952	1.5839	0.844478	0.67
Cutting Plane	100	NA	0.0630	1.0951	1.5841	0.844477	0.91
Analytical Int.	1000	NA	0.0680	1.0108	1.7322	1.049276	174.24
Numerical Int.	1000	1000	0.0680	1.0108	1.7322	1.049276	1.21
Subgradient	1000	NA	0.0680	1.0109	1.7321	1.049276	0.34
Coord. Descent	1000	NA	0.0680	1.0109	1.7321	1.049276	1.20
Cutting Plane	1000	NA	0.0680	1.0109	1.7320	1.049276	2.07
Analytical Int.	10000	—	—	—	—	—	—
Numerical Int.	10000	1000	0.0818	1.0048	1.6430	1.092066	5.27
Subgradient	10000	NA	0.0818	1.0048	1.6429	1.092033	2.88
Coord. Descent	10000	NA	0.0834	1.0047	1.6378	1.092040	0.88
Cutting Plane	10000	NA	0.0817	1.0049	1.6432	1.092033	3.00

Table 4. Example A: Solution vectors and computing times (sec.) for the superquantile regression problem with varying computational methods, and sample sizes ν.

Dual Method	α	ν	c_0	c_1	c_2	Function Value	Time
Subgradient Method	0.25	100	-0.0478	1.1038	0.6420	0.502796	0.14
		1000	-0.0557	0.9988	0.6726	0.584710	0.33
		10000	-0.0398	0.9990	0.6048	0.608587	2.61
	0.50	100	0.0163	1.0901	0.8440	0.598695	0.14
		1000	0.0309	0.9943	0.8822	0.705208	0.33
		10000	0.0390	1.0014	0.8239	0.732942	2.99
	0.75	100	0.0390	1.1029	1.2056	0.729180	0.14
		1000	0.0762	0.9976	1.2239	0.875049	0.33
		10000	0.0742	1.0009	1.2050	0.905114	2.93

Table 5. Example A: Solution vectors and computing times (sec.) for solving D^ν when implementing the subgradient method with varying probability levels α and number of observations ν.

third decimal places, which makes the magnitude of such differences negligible.

Table 6 presents the solution vectors and computing times for the coordinate descent method for different probability levels α and sample sizes ν. Similarly to the subgradient method, we realize that only the sample size ν has a significant effect on the computing times.

Dual Method	α	ν	c_0	c_1	c_2	Function Value	Time
		100	-0.0478	1.1038	0.6419	0.502796	0.62
	0.25	1000	-0.0227	0.9995	0.5937	0.585203	0.16
Coordinate		10000	-0.0392	0.9990	0.6034	0.608588	2.51
		100	0.0163	1.0901	0.8439	0.598695	0.65
Descent	0.50	1000	0.0340	0.9943	0.8734	0.705218	0.21
		10000	0.0403	1.0014	0.8205	0.732944	1.58
Method		100	0.0390	1.1029	1.2056	0.729180	0.70
	0.75	1000	0.0771	0.9977	1.2213	0.875050	0.21
		10000	0.0763	1.0009	1.1987	0.905121	1.03

Table 6. Example A: Solution vectors and computing times (sec.) for solving D^ν when implementing the coordinate descent method with varying probability levels α and number of observations ν.

Dual Method	α	ν	c_0	c_1	c_2	Function Value	Time
		100	-0.0479	1.1038	0.6420	0.502797	1.13
	0.25	1000	-0.0556	0.9989	0.6723	0.584710	1.51
Cutting		10000	-0.0399	0.9989	0.6049	0.608587	2.90
		100	0.0162	1.0901	0.8440	0.598695	2.23
Plane	0.50	1000	0.0307	0.9944	0.8827	0.705208	1.21
		10000	0.0389	1.0017	0.8243	0.732943	1.22
Method		100	0.0390	1.1029	1.2056	0.729180	0.80
	0.75	1000	0.0763	0.9976	1.2238	0.875049	1.97
		10000	0.0739	1.0008	1.2059	0.905114	1.16

Table 7. Example A: Solution vectors and computing times (sec.) for solving D^ν when implementing the cutting plane method with varying probability levels α and number of observations ν.

A different result is obtained when we implement the cutting plane method, as shown in Table 7. The computing time differences are not significant in any of the cases. Here we run the cutting plane method with bounds on the vectors (c_1^k, c_2^k), for iteration k. Decreasing these bounds by making them more restrictive, and reducing the maximum number of cuts that the algorithm can add, reduces the computing times shown by Column 8 in Table 7, and the magnitudes of the computing times differences become even less significant.

Out of curiosity, if we implement the subgradient method for this example with ten times more iterations, i.e., a total of 10000 iterations, and reduce the stepsize to $\delta = 0.01$, we find that the solution vectors are exactly the same as the ones presented in Table 5, with the same objective function values, but the computing times increase by at least a factor of 10 in the cases of $\nu = 10000$ and $\nu = 100000$, a factor of 18 for $\nu = 1000$, and a factor of 30 for $\nu = 100$. This shows how important the selection of the right stepsize and maximum number of iterations is.

From this example we conclude that for small sample sizes it is beneficial to run the primal method using analytical integration, since we obtain the exact solution vector and the computing times are not drastically higher than solving D^ν. As the sample size increases, the results show that we should rely on the dual method and implement the cutting plane method or any other algorithm that is comparable to the cutting plane method. Another aspect we observe is the fact that the probability level α does not produce any visible effect on the dual methods computing times. To the contrary, the primal methods, with analytical or numerical integration schemes, are clearly affected due to changes in α since the integral interval is adjusted accordingly and the number of variables changes. Implementing the primal methods with numerical integration schemes implies the wise selection of the number of subintervals μ according to the sample size ν and probability level α.

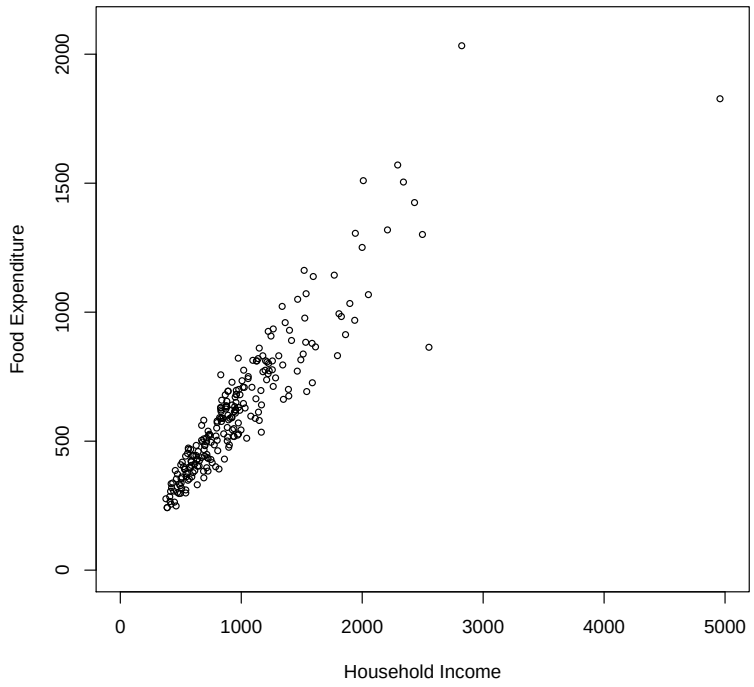
B. ENGEL DATA

This next example is based on a data set originally worked by Ernst Engel in 1857, and used by Koenker and Bassett in their regression quantiles studies (Koenker & Bassett Jr., 1982). Engel presents this data set to show his hypothesis that the annual expenditures on food for Belgian working class families increase less than the increase of their annual household incomes. In Koenker (2005), the author uses this data set as an example to address the issues of estimating the asymptotic covariance matrix in statistical inference for quantile regression and estimates six quantile regression functions for probability levels $\alpha \in \{0.05, 0.10, 0.25, 0.75, 0.90, 0.95\}$. For this example, we are interested in comparing these obtained quantile regression functions with superquantile regression functions at the same probability levels α, and verify how both regression techniques differ conceptually.

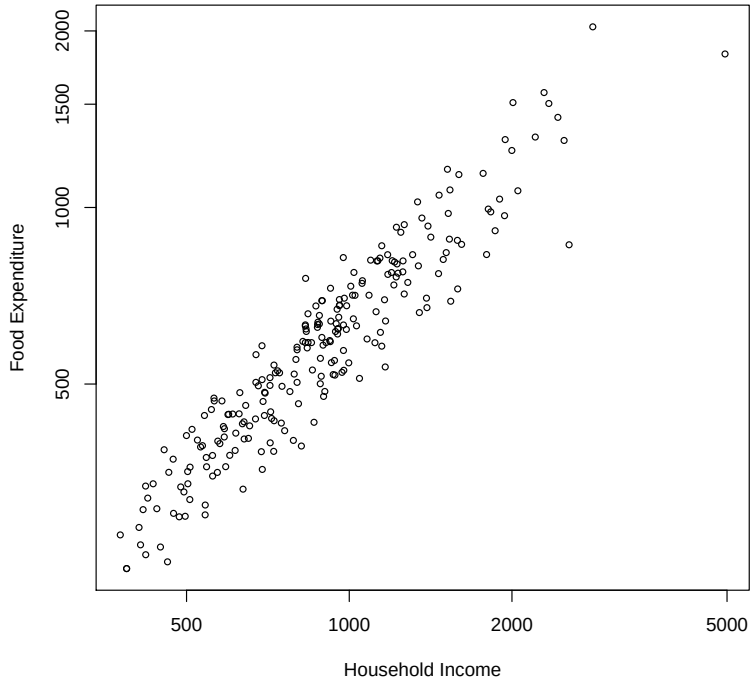
We have a data set of 235 observations of the income and the expenditure on food in Belgian francs for Belgian working class annual households, see Figure 8(a). As done in Koenker (2005), we also transform both variables to the logarithmic scale, see Figure 8(b). We seek to quantify the food expenditure Y and consider a linear regression function $f_1(X) = c_0 + cX$, where X is the explanatory random variable that represents the household income for Belgian working class.

In Figure 9(b), we observe the α-quantile regression models, for probability levels $\alpha \in \{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$, in logarithmic scale. Here we also include the least squares and the 0.50-quantile regression functions for comparison and highlight the obtained 0.75-quantile regression function that we use later in this example. Although some of these quantile regression functions look parallel, their slopes are distinct; see Koenker (2005) for further discussion. These slope differences are more evident in Figure 9(a).

In Figure 10, we present the α-superquantile regression models, for different probability levels $\alpha \in \{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$. Again we include the least squares and the 0.50-superquantile regression functions and highlight the ob-

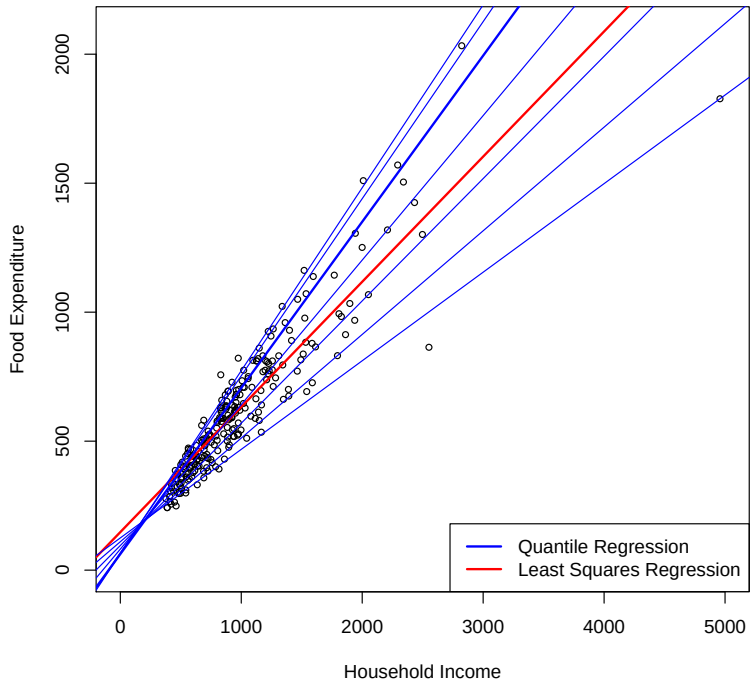


(a) Original display.

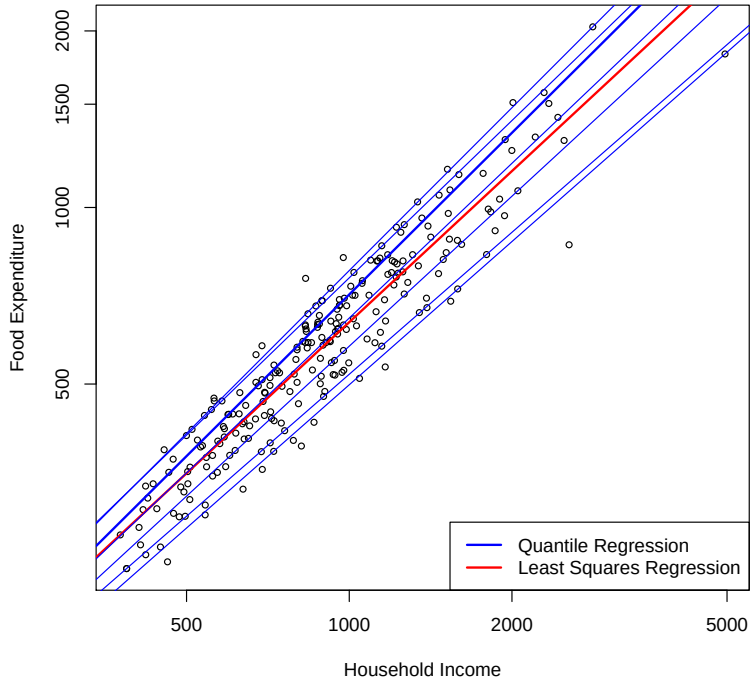


(b) Logarithmic scale display.

Figure 8. Example B: Engel data set.



(a) Original display.



(b) Logarithmic scale display.

Figure 9. Example B: Least squares and quantile regression functions, for varying α.

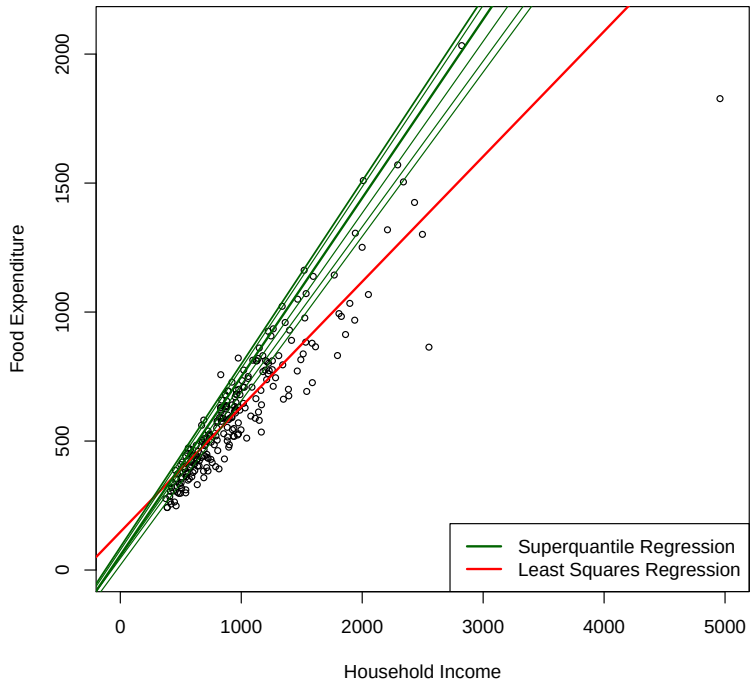
tained 0.75-superquantile regression fit.

An interesting detail shown in Figure 9 is that the obtained quantile regression models nearly "span" the observations, i.e., we have regression functions above and below the least squares regression fit. As we observe in Figure 10, the superquantile regression models for varying probability levels α do not have such property. One would have to change the orientation of the original problem in order to obtain regression functions below the least squares regression model, since $\bar{q}_0(Y) = E[Y]$.

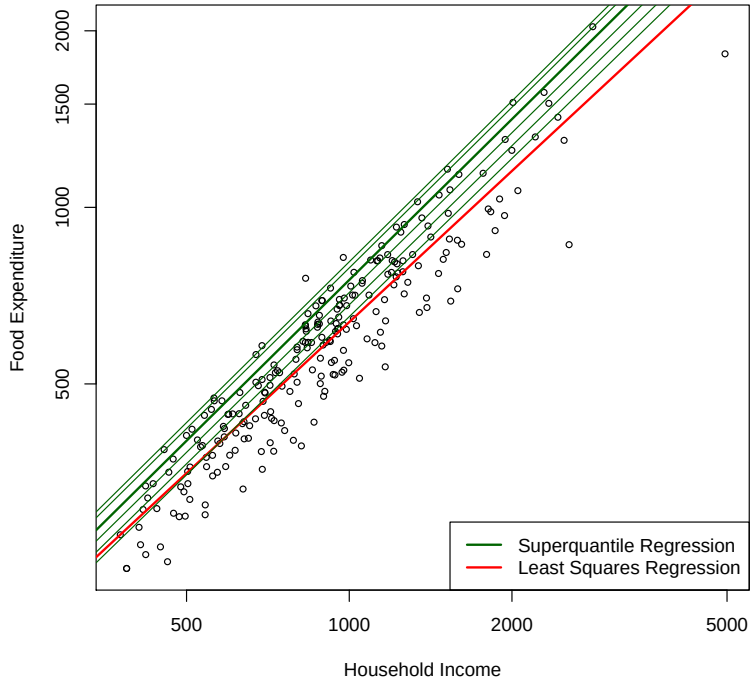
In order to compare the obtained regression vectors and the corresponding coefficient of determination for the model $f_1(x) = c_0 + c_1 x$, we refer to Table 8. We cosider the same probability levels α as shown in Figures 9 and 10. Due to the small sample size, 235 observations, we solve the deviation-based superquantile regression problem by analytical integration, P^{ν}_{LP}. We refer to Figure 11(a) to show how close the quantile and superquantile linear regression functions are in the case where $\alpha = 0.75$.

We now consider a quadratic model of the form $f_2(x) = c_0 + c_1 x + c_2 x^2$. In Figure 11(b), we observe the different quadratic fits for least squares, quantile, and superquantile regressions. Although both graphs in Figure 11 show that the 0.75-superquantile regression functions look exactly alike, Figure 11(b) actually has a curvature that can be noted using different scales on the horizontal axis. Table 8 shows the obtained regression vectors (c_0, c_1, c_2) for the quadratic model, using distinct regression techniques. We note that the coefficient of determination for the linear are slightly smaller than for the quadratic models, which means that adding the term $c_2 x^2$ slightly improves the obtained results in some sense.

In Figure 12, we visualize quantile and superquantile regression functions for varying probability levels α. It is interesting to notice how the quantile regression fits are severely influenced by one observation where four quantile regression functions cross each other just below the least squares fit, represented by the big black dot. To the contrary, the obtained superquantile regression fits are not greatly influenced by this observation and depict other observations.



(a) Original display.

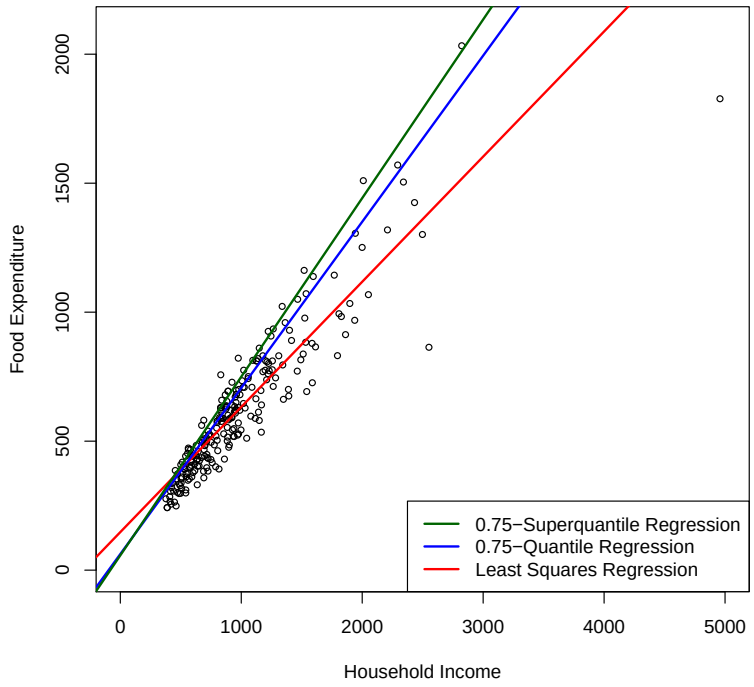


(b) Logarithmic scale display.

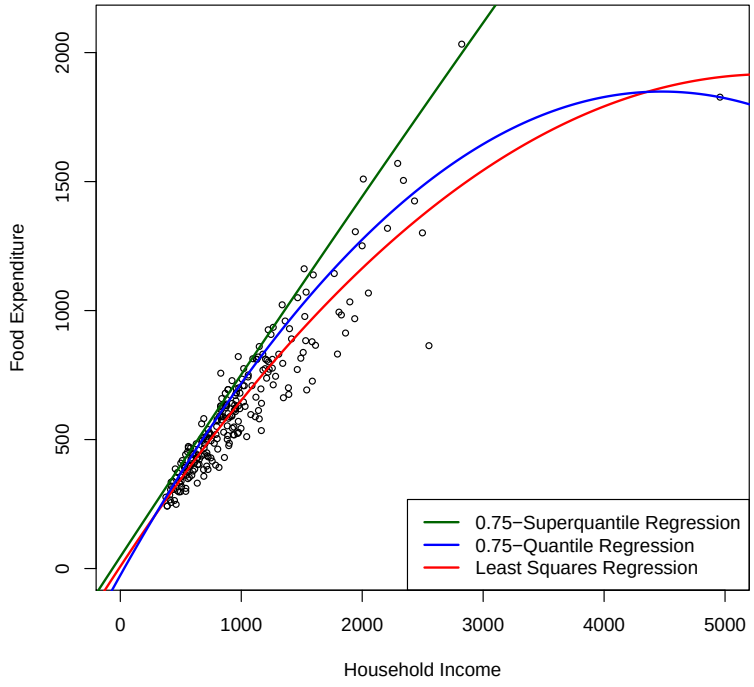
Figure 10. Example B: Least squares and superquantile regression functions, for varying α.

Regression Model	α	c_0	c_1	c_2	$\bar{R}^2_\alpha$
Least Squares	NA	147.475	0.4852	—	0.8304
Quantile	0.05	124.880	0.3434	—	—
	0.10	110.142	0.4018	—	—
	0.25	95.4835	0.4741	—	—
	0.50	81.4822	0.5602	—	—
	0.75	62.3966	0.6440	—	—
	0.90	67.3509	0.6863	—	—
	0.95	64.1040	0.7091	—	—
Superquantile	0.05	18.8791	0.6370	—	0.6882
	0.10	27.0860	0.6387	—	0.6913
	0.25	45.2404	0.6425	—	0.7043
	0.50	52.3684	0.6657	—	0.7322
	0.75	57.3732	0.6924	—	0.7716
	0.90	77.4796	0.7039	—	0.8070
	0.95	88.6620	0.7097	—	0.8223
Least Squares	NA	8.0060	0.7100	-6.603e-5	0.8671
Quantile	0.05	-31.7001	0.6815	-1.295e-4	—
	0.10	52.6260	0.5009	-2.884e-5	—
	0.25	22.8226	0.6123	-5.009e-5	—
	0.50	5.7593	0.7243	-7.198e-5	—
	0.75	-26.0488	0.8378	-9.360e-5	—
	0.90	72.2423	0.6724	7.838e-6	—
	0.95	44.3764	0.7445	-1.419e-5	—
Superquantile	0.05	-28.7584	0.7354	-4.243e-5	0.6903
	0.10	-13.3480	0.7212	-3.498e-5	0.6928
	0.25	17.2230	0.6946	-1.896e-5	0.7050
	0.50	32.8155	0.7034	-1.439e-5	0.7327
	0.75	45.6962	0.7144	-8.130e-6	0.7717
	0.90	54.6966	0.7461	-1.467e-5	0.8079
	0.95	53.0274	0.7777	-2.522e-5	0.8241

Table 8. Example B: Solution vectors (c_0, c_1) and coefficients of determination for the linear model of the form $f_1(x) = c_0 + c_1 x$, and solution vectors (c_0, c_1, c_2) and coefficients of determination for the quadratic model of the form $f_2(x) = c_0 + c_1 x + c_2 x^2$.

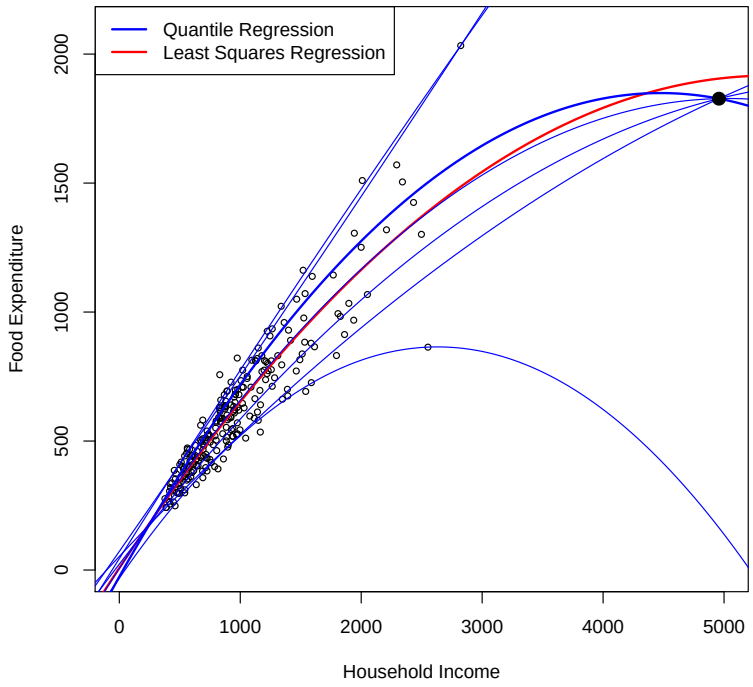


(a) Linear model $f_1(x) = c_0 + c_1 x$.

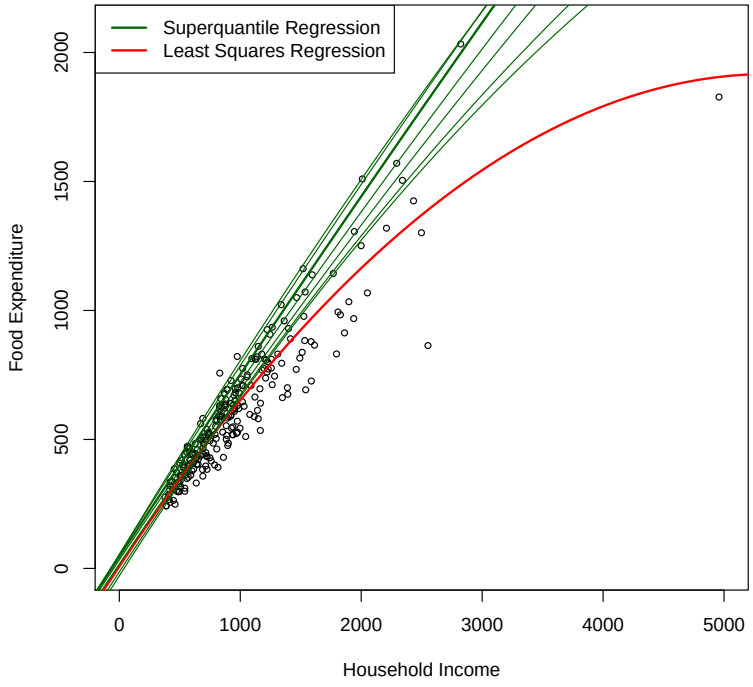


(b) Quadratic model $f_2(x) = c_0 + c_1 x + c_2 x^2$.

Figure 11. Example B: Regression functions for linear and quadratic models.



(a) Quantile regression for varying probability levels α.



(b) Superquantile regression for varying probability levels α.

Figure 12. Example B: Least squares, quantile, and superquantile regression functions for the quadratic model $f_2(x) = c_0 + c_1 x + c_2 x^2$.

As a conclusion to this example, we note that superquantile regression brings additional information concerning the tail realizations of our loss random variable. The linear fits from quantile and superquantile regressions are close, with only a slight difference in slope. However, the quadratic superquantile model provides a distinct perspective. In the quadratic case, quantile regression is highly affected in a dubious manner by one observation.

C. BROWNLEE STACK LOSS PLANT DATA

This example is based on a data set with 21 observations from the Brownlee stack loss plant data set, which defines the oxidation of ammonia (NH3) to nitric acid (HNO3) of a plant, as described in detail in Brownlee (1965).

We seek to estimate the stack loss random variable Y, representing ten times the percentage of ammonia going into the plant that escapes from the absorption tower, using three explanatory random variables: air flow (X_{af}), which represents the rate of operation of the plant; water temperature (X_{wt}), which denotes the temperature of cooling water circulated through coils in the absorption tower; and acid concentration (X_{ac}), [per 1000, minus 500].

Figure 13 shows the scatterplot matrix of the stack loss data, where we observe the pairwise correlations. Here we notice a linear correlation between stack loss and air flow and a polynomial correlation between stack loss and water temperature. We explore these two possible models and compare the obtained results with coefficient of determination calculations, as described in Section II.C.

We first consider a linear model of the form $f_1(x) = c_0 + c_{\text{af}}x_{\text{af}} + c_{\text{wt}}x_{\text{wt}} + c_{\text{ac}}x_{\text{ac}}$. Table 9 shows the obtained regression coefficients after solving P^{ν}_{LP}. All the instances of problem P^{ν}_{LP} take approximately one quarter of a second to run due to the small number of observations in the data sample.

From Table 9, we conclude that a linear model with all three explanatory random variables is reasonable. It is interesting to note that the resulting coefficients of

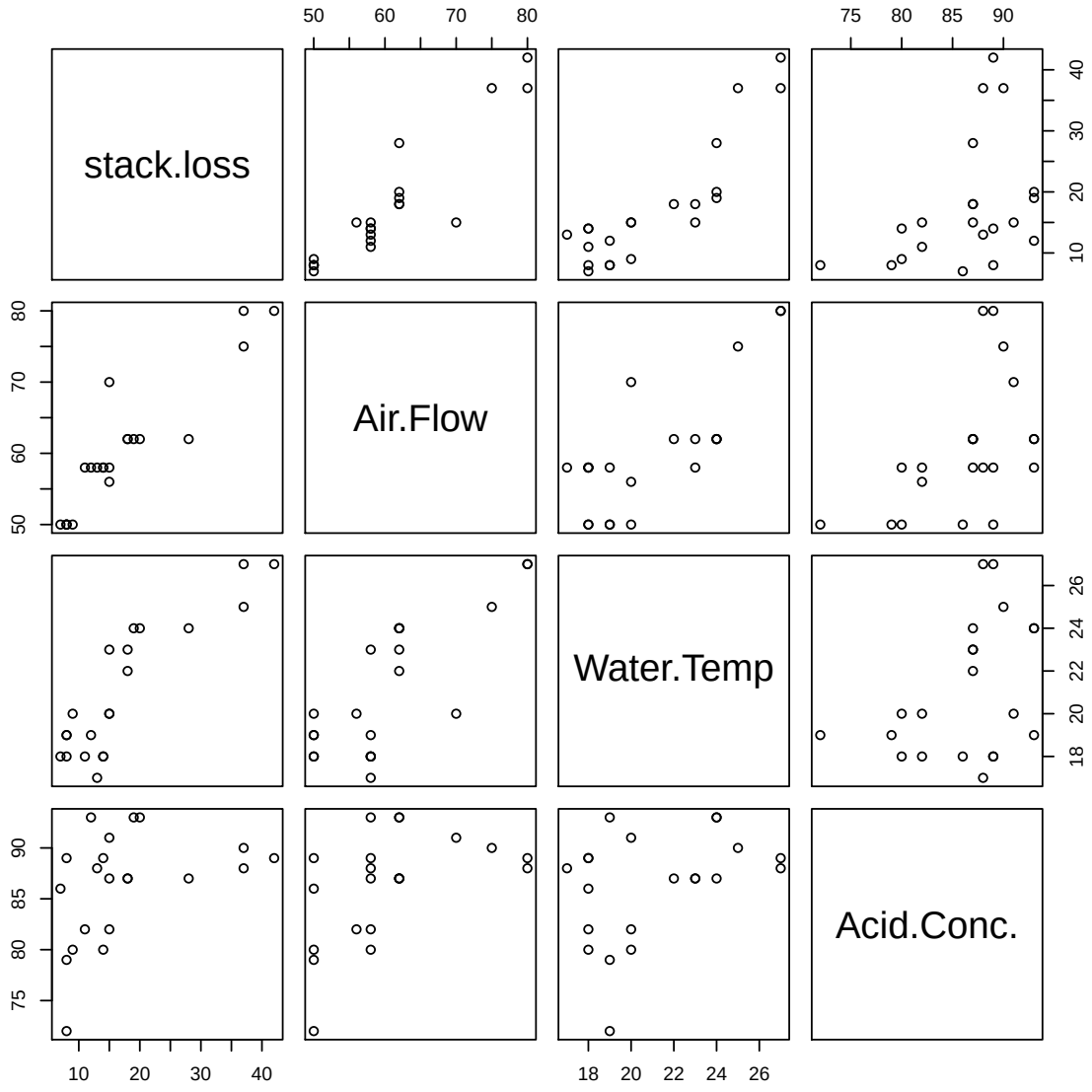


Figure 13. Example C: Stack loss data scatterplot matrix.

Regression	α	c_0	c_{af}	c_{wt}	c_{ac}	$\bar{R}^2_\alpha$	$\bar{R}^2_{\alpha,\text{Adj}}$
Least Squares	NA	-39.9197	0.7156	1.2953	-0.1521	0.9136	0.8983
Quantile	0.25	-36.0000	0.5000	1.0000	0.0000	—	—
	0.50	-39.6899	0.8319	0.5739	-0.0609	—	—
	0.75	-54.1897	0.8707	0.9828	0.0000	—	—
	0.90	-58.5433	0.7930	1.3054	0.0382	—	—
Superquantile	0.25	-55.1432	0.8056	1.2037	0.0000	0.7478	0.7033
	0.50	-58.6210	0.7930	1.3054	0.0382	0.7750	0.7353
	0.75	-60.1368	0.7500	1.4561	0.0570	0.8050	0.7706
	0.90	-58.4620	0.5246	1.8584	0.1073	0.8231	0.7919

Table 9. Example C: Regression vectors, $\bar{R}^2_\alpha$, and $\bar{R}^2_{\alpha,\text{Adj}}$ for the linear model f_1 which includes all explanatory variables, and for different probability levels α.

α	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
$\bar{R}^2_\alpha$	0.7384	0.7402	0.7423	0.7447	0.7478	0.7516	0.7563	0.7618	0.7682
α	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
$\bar{R}^2_\alpha$	0.7750	0.7818	0.7883	0.7944	0.8001	0.8050	0.8110	0.8173	0.8231

Table 10. Example C: Coefficients of determination for different probability levels α.

determination $\bar{R}^2_\alpha$ and adjusted coefficients of determination $\bar{R}^2_{\alpha,\text{Adj}}$ for superquantile regression increase with α, which lead us to further experiment for various probability levels α. Table 10 shows the obtained coefficients of determination for varying α.

We next analyze a simpler model, using water temperature as the only available explanatory variable, and compare the corresponding linear $f_2(x) = c_0 + c_{\text{wt}}x_{\text{wt}}$ and quadratic models $f_3(x) = c_0 + c_{\text{wt}}x_{\text{wt}} + c_{\text{wt2}}x^2_{\text{wt}}$; see Table 11. For the situation where one only has water temperature as the explanatory variable, applying the quadratic model f_3 slightly reduces the coefficients of determination. However we plot the obtained regression functions, see Figure 14(b), and notice that the quadratic models better represent the data.

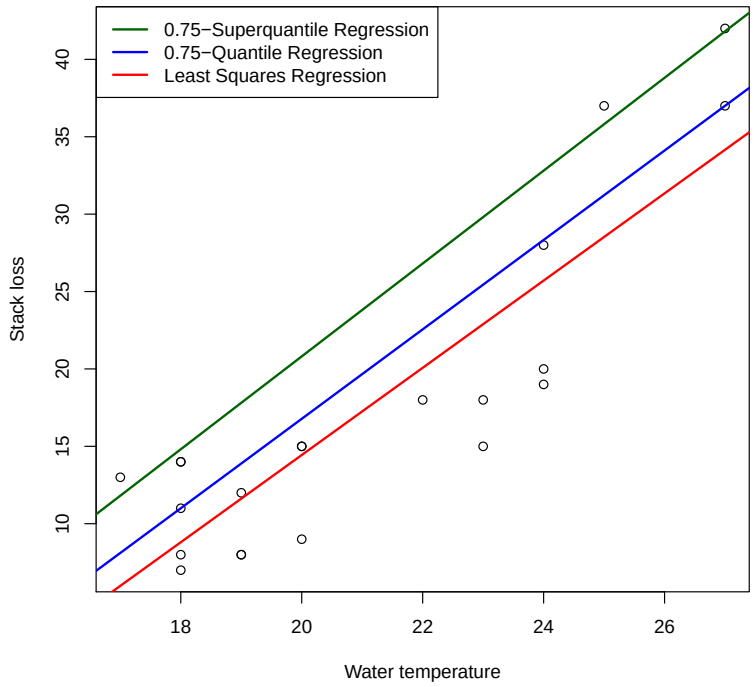
It is interesting to notice that the 0.90-quantile and 0.90-superquantile regression functions are exactly the same in the quadratic model f_3. This is due to a small

Model	Regression	α	c_0	c_{wt}	c_{wt2}	$\bar{R}^2_\alpha$	$\bar{R}^2_{\alpha,\text{Adj}}$
f_2	Least Squares	NA	-41.9109	2.8174	—	0.7665	0.7542
	Quantile	0.25	-32.0000	2.1667	—	—	—
		0.50	-47.8571	3.1429	—	—	—
		0.75	-41.0000	2.8889	—	—	—
		0.90	-42.0000	3.1111	—	—	—
	Superquantile	0.25	-43.6667	3.0000	—	0.5649	0.5420
		0.50	-41.7619	3.0000	—	0.5954	0.5741
		0.75	-39.1905	3.0000	—	0.6440	0.6250
		0.90	-38.0476	3.0000	—	0.6715	0.6540
f_3	Least Squares	NA	151.5654	-15.2555	0.4131	0.8755	0.8617
	Quantile	0.25	148.6000	-15.1583	0.4083	—	—
		0.50	200.8500	-19.8333	0.5167	—	—
		0.75	110.1429	-11.1381	0.3191	—	—
		0.90	205.5714	-20.6714	0.5571	—	—
	Superquantile	0.25	167.5589	-16.9167	0.4583	0.6676	0.6306
		0.50	183.9524	-18.5000	0.5000	0.6884	0.6538
		0.75	205.4789	-20.6714	0.5571	0.7490	0.7211
		0.90	205.5714	-20.6714	0.5571	0.7792	0.7546

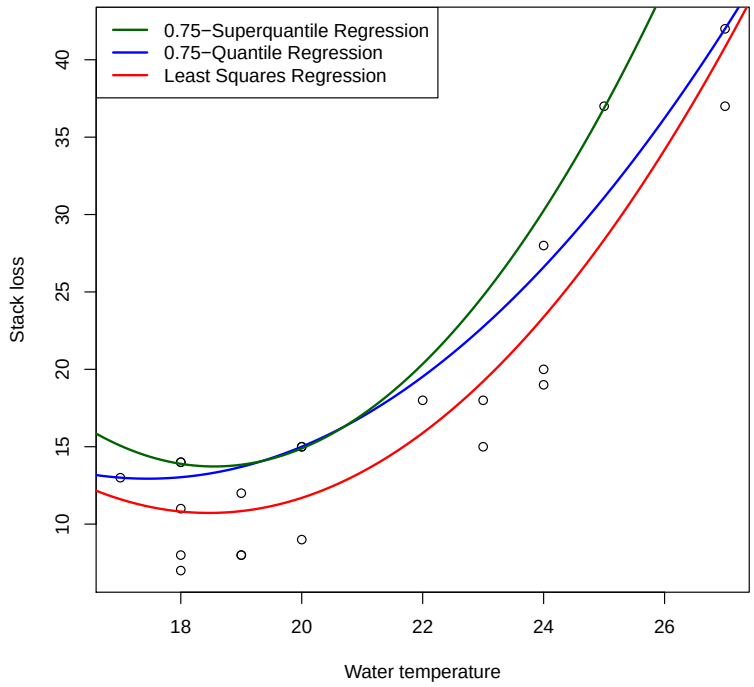
Table 11. Example C: Regression vectors, $\bar{R}^2_\alpha$, and $\bar{R}^2_{\alpha,\text{Adj}}$ for linear and quadratic models, f_2 and f_3, respectively, for varying probability levels α.

data set and how the observations are dispersed. For example, here we have three observations at sample point $(x^j, y^j) = (20, 15)$. For such a very small data set, having coincident observations does not help obtaining better quantile or superquantile regression fits. We notice that the 0.75-quantile regression function is a clear example of how having small data sets aggravated by overlapping observations influences the obtained regression vector and may cause the function to shift accordingly. In this case, we realize that both 0.75-quantile and 0.75-superquantile regression functions cross the point $(x^j, y^j) = (20, 15)$.

As a conclusion to this example, we note that small data sets in superquantile regression are problematic to deal with. As a thumb rule, one needs $1/(1-\alpha)$ times more observations when performing superquantile regression than when in the case of least squares regression. Therefore the obtained approximating regression vectors



(a) Linear model $f_2(x) = c_0 + c_{wt}x_{wt}$.



(b) Quadratic model $f_3(x) = c_0 + c_{wt}x_{wt} + c_{wt2}x_{wt}^2$.

Figure 14. Example C: Regression functions for linear and quadratic models.

for small data sets should be considered with care when used in decision making processes.

D. INVESTMENT ANALYSIS

The next example is a case study taken from the "Style Classification with Quantile Regression" documentation in Portfolio Safeguard, by American Optimal Decisions, Inc. (2011), and deals with the negative return of the Fidelity Magellan Fund as predicted by the explanatory variables Russell 1000 Growth Index (X_{RLG}), Russell 1000 Value Index (X_{RLV}), Russell Value Index (X_{RUJ}), and Russell 2000 Growth Index (X_{RUO}). We change the orientation from "return" to "negative return" to be consistent with the orientation of a loss random variable in this dissertation. The indices classify the style of the fund; see American Optimal Decisions, Inc. (2011) for details. There are $\nu = 1264$ total observations available.

We start by considering a linear model $f_1(x) = c_0 + c_{\mathrm{RLG}}x_{\mathrm{RLG}} + c_{\mathrm{RLV}}x_{\mathrm{RLV}} + c_{\mathrm{RUJ}}x_{\mathrm{RUJ}} + c_{\mathrm{RUO}}x_{\mathrm{RUO}}$ and compare the obtained approximate regression vectors for least squares, quantile, and superquantile regression models under a probability level α =0.75 and 0.90, as shown in Rows 2-6 of Table 12. $DSqR^{\nu}$ is solved through $P^{\nu,\mu}_{\mathrm{Num}}$ with Simpson's rule as the integration scheme and $\mu = 1000$ subintervals, while quantile regression is carried out directly in Portfolio Safeguard Shell Environment (American Optimal Decisions, Inc., 2011). Table 12 (last column) also shows the corresponding adjusted coefficients of determination. The fits are good and a majority of the variability in the data is captured. However, the small values of c_{RUO} and also the corresponding p-value from the least squares regression point to the possible merit of dropping X_{RUO} as explanatory variable. We from now on focus on superquantile regression. A new model $f_2(x) = c_0 + c_{\mathrm{RLG}}x_{\mathrm{RLG}} + c_{\mathrm{RLV}}x_{\mathrm{RLV}} + c_{\mathrm{RUJ}}x_{\mathrm{RUJ}}$ yields the approximate regression vectors of Table 12 (Rows 7-8), which also shows the obtained adjusted coefficients of determination $\bar{R}^2_{\alpha,\mathrm{Adj}}$. The fact that we analyze $\bar{R}^2_{\alpha,\mathrm{Adj}}$ instead of $\bar{R}^2_{\alpha}$ enable us to better compare fits across models with different numbers

Model	Regression	α	c_0	c_{RLG}	c_{RLV}	c_{RUJ}	c_{RUO}	$\bar{R}^2_{\alpha,\mathrm{Adj}}$
f_1	LS	NA	0.0010	-0.5089	-0.5180	0.0484	0.0061	0.9823
	Quantile	0.75	0.0045	-0.5438	-0.4518	0.0159	0.0173	—
		0.90	0.0089	-0.5177	-0.4602	0.0156	-0.0001	—
	Super-quantile	0.75	0.0095	-0.5036	-0.4723	0.0192	0.0009	0.8731
		0.90	0.0138	-0.4837	-0.4912	0.0223	-0.0019	0.8718
f_2	Super-quantile	0.75	0.0095	-0.5028	-0.4728	0.0200	—	0.8733
		0.90	0.0138	-0.4855	-0.4906	0.0210	—	0.8720
f_3	Super-quantile	0.75	0.0137	-0.8228	—	—	—	0.7380
		0.90	0.0218	-0.8189	—	—	—	0.7248
		0.75	0.0321	—	-1.0668	—	—	0.5940
		0.90	0.0475	—	-1.0727	—	—	0.5702
		0.75	0.0515	—	—	-0.7745	—	0.4103
		0.90	0.0714	—	—	-0.6949	—	0.4162
		0.75	0.0344	—	—	—	-0.5498	0.3962
		0.90	0.0512	—	—	—	-0.5145	0.2593

Table 12. Example D: Approximate least squares (LS), quantile, and superquantile regression vectors and $\bar{R}^2_{\alpha,\mathrm{Adj}}$ for models f_1, f_2, and f_3.

of explanatory variables. In comparison, the fit improves slightly by dropping X_{RUO}.

We further reduce the model to a single explanatory variable, $f_3(x) = c_0 + c_i x_i$, with $i \in \{\mathrm{RLG}, \mathrm{RLV}, \mathrm{RUJ}, \mathrm{RUO}\}$, and examine the four possibilities in Rows 9-16 of Table 12. We find that $\bar{R}^2_{\alpha,\mathrm{Adj}}$ deteriorates, but only moderately for the model $c_0 + c_{\mathrm{RLG}} X_{\mathrm{RLG}}$. This simple model captures much of the variability in the data set. A somewhat poorer fit is achieved by X_{RLV}, which is illustrated in Figure 15, for $\alpha = 0.90$. That figure also depicts the corresponding quantile and least squares regression lines. It is apparent that superquantile regression provides a distinct perspective from the other regression techniques of potentially significant value to a decision maker.

E. U.S. NAVY HELICOPTER PILOTS DATA

This example considers the results of an online survey of winged Naval helicopter pilots of the U.S. Navy; see Phillips (2011) for details. Her goal is to verify if helicopter pilots back pain is a concern among the helicopter community and to define this problem's implications. Although this is an important and real issue in

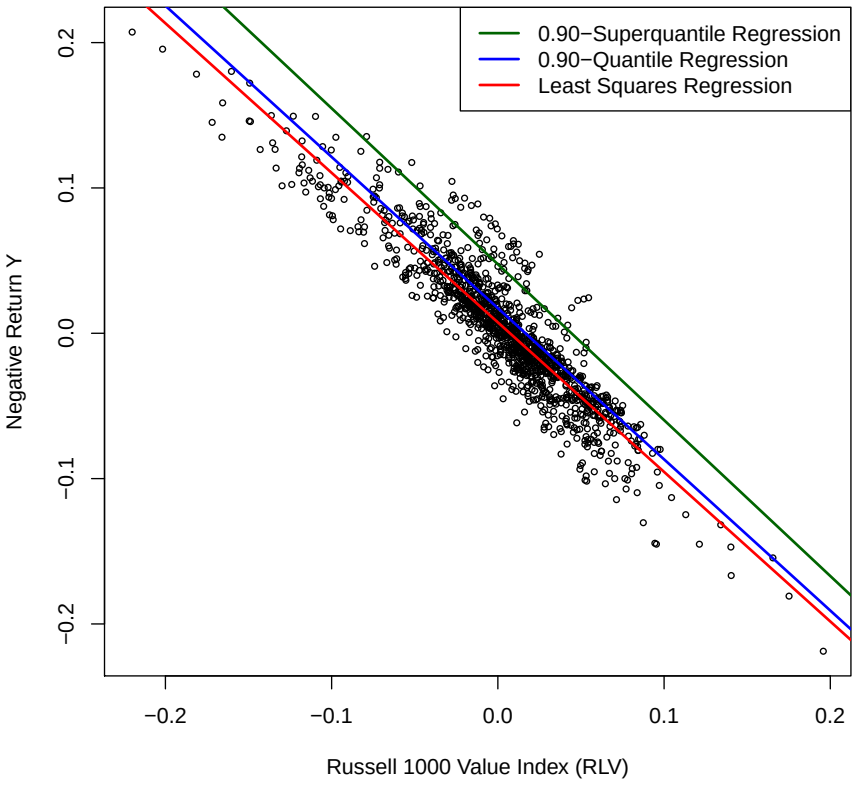


Figure 15. Example D: Regression lines for model $c_0 + c_{\text{RLV}} X_{\text{RLV}}$.

the helicopter community, we do not use the superquantile regression technique developed in this dissertation to estimate helicopter pilots' back pain frequency due to the categorical nature of this random variable. Instead we utilize the available data set to estimate the total flight hours Y for naval helicopter pilots. As explanatory variables we have the number of years a helicopter pilot has flown for the U.S. Navy (X_{years}), and their body mass index (X_{BMI}), available through a formula derived using the available data on height and weight of helicopter pilots.

Since we only consider those pilots that answered questions in the "Demographics" and "Flight Hour Info" sections, see Appendix A in Phillips (2011), of the 648 pilots that completed the survey, we only use 633 observations. Figure 16 displays these observations in a pairwise scatterplot matrix. As expected, one clearly depicts the linear correlation between years an helicopter pilot has flown for the U.S. Navy

and the estimated total number of flight hours.

We first consider a regression function of the form $f(x) = c_0 + c_{\text{years}}x_{\text{years}} + c_{\text{BMI}}x_{\text{BMI}}$ and vary the probability levels α. Rows 2-10 in Table 13 report the obtained

Model	Regression	α	c_0	c_{years}	c_{BMI}	$\bar{R}^2_\alpha$	$\bar{R}^2_{\alpha,\text{Adj}}$
f_1	Least Squares	NA	51.70	161.22	0.9176	0.7780	0.7773
	Quantile	0.25	-48.71	146.67	0.3418	—	—
		0.50	-55.56	177.78	0.0000	—	—
		0.75	0.0000	200.00	0.0000	—	—
		0.99	1233.3	322.49	-46.565	—	—
	Superquantile	0.25	-47.03	200.00	0.000	0.6094	0.6081
		0.50	2.1827	208.69	-0.1809	0.6205	0.6193
		0.75	116.71	223.21	-2.6097	0.6147	0.6134
		0.99	244.33	323.79	-75.903	0.4754	0.4738
f_2	Least Squares	NA	74.84	161.30	—	0.7780	0.7776
	Quantile	0.25	-40.00	146.67	—	—	—
		0.50	-55.56	177.78	—	—	—
		0.75	0.0000	200.00	—	—	—
		0.99	-93.75	343.75	—	—	—
	Superquantile	0.25	-47.03	200.00	—	0.6094	0.6088
		0.50	-1.781	208.57	—	0.6205	0.6199
		0.75	49.721	223.13	—	0.6146	0.6140
		0.99	247.55	350.00	—	0.4538	0.4529

Table 13. Example E: Regression vectors, $\bar{R}^2_\alpha$, and $\bar{R}^2_{\alpha,\text{Adj}}$ for model $f_1(x) = c_0 + c_{\text{years}}x_{\text{years}} + c_{\text{BMI}}x_{\text{BMI}}$ and $f_2(x) = c_0 + c_{\text{years}}x_{\text{years}}$ at varying probability levels α.

solution vectors for model f_1, the corresponding coefficients of determination $\bar{R}^2_\alpha$, and adjusted coefficients of determination $\bar{R}^2_{\alpha,\text{Adj}}$. The fits are reasonable but the p-value for c_{BMI} from the least squares regression suggests the possible benefit of dropping X_{BMI} as explanatory variable. With this in mind, we drop the explanatory random variable X_{BMI} from our new model. Before we move on to the next model, we notice that the obtained 0.99-quantile regression solution vector is correct although its intercept looks way larger compared to other approximate solution vectors.

Second we consider a single-variable model of the form $f_2(x) = c_0 + c_{\text{years}}x_{\text{years}}$,

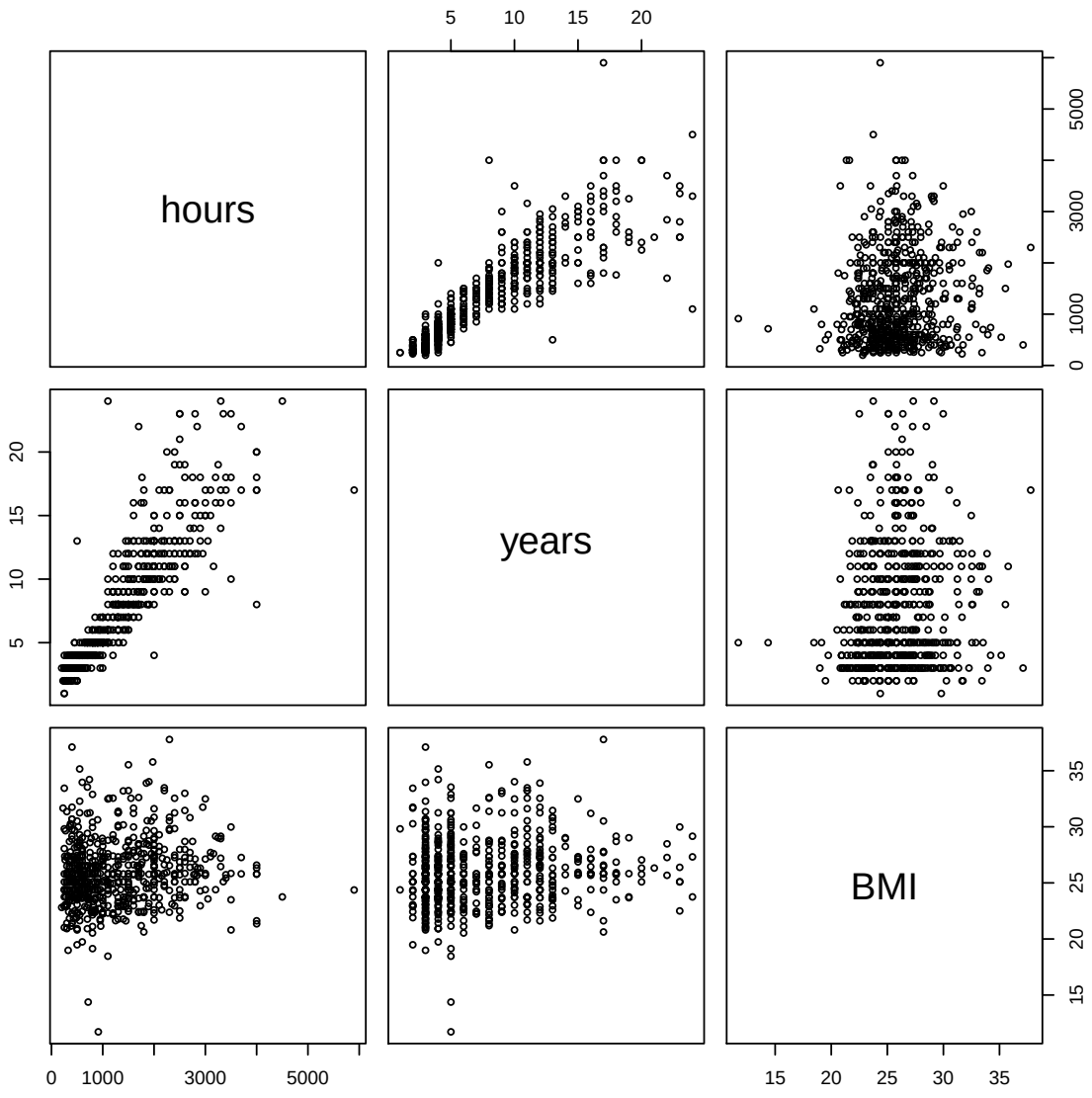


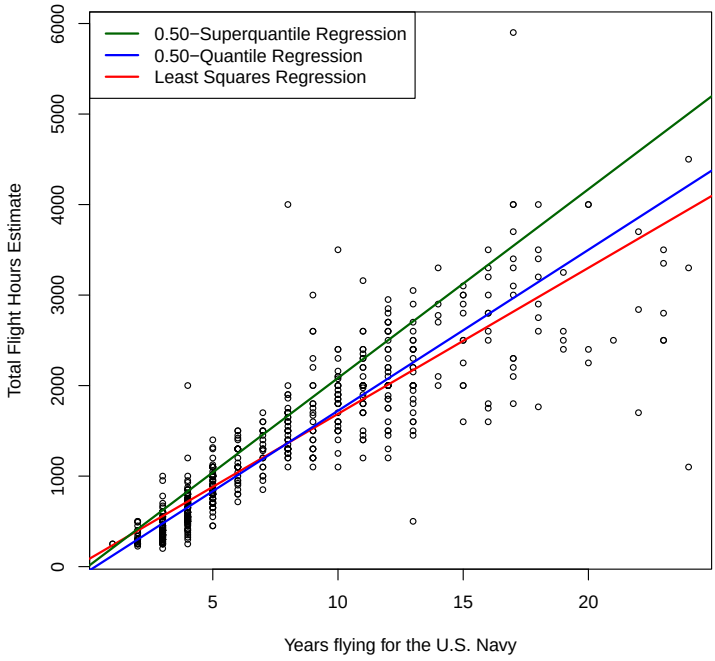
Figure 16. Example E: U.S. Navy helicopter pilots data scatterplot matrix.

and obtain the results presented in Rows 11-19 of the same Table 13. The adjusted coefficients of determination $\bar{R}^2_{\alpha,\text{Adj}}$ slightly increase in the cases of least squares and superquantile regressions techniques, where $\alpha \in \{0.25, 0.50, 0.75\}$. Figure 17(a) shows the corresponding regression lines for the linear model f_2, at a fixed probability level $\alpha = 0.50$. It is interesting to notice that the quantile regression line for $\alpha = 0.50$ has a negative intercept, while the least squares and superquantile regression functions intercept the y-axis at higher values. Another aspect we learn from Figure 17(a) is the importance of the magnitude of errors in regression. This is evident when we compare both quantile and superquantile regression lines. Superquantile regression responds to the observations that have larger errors, emphasizing those observations that we might consider outliers.

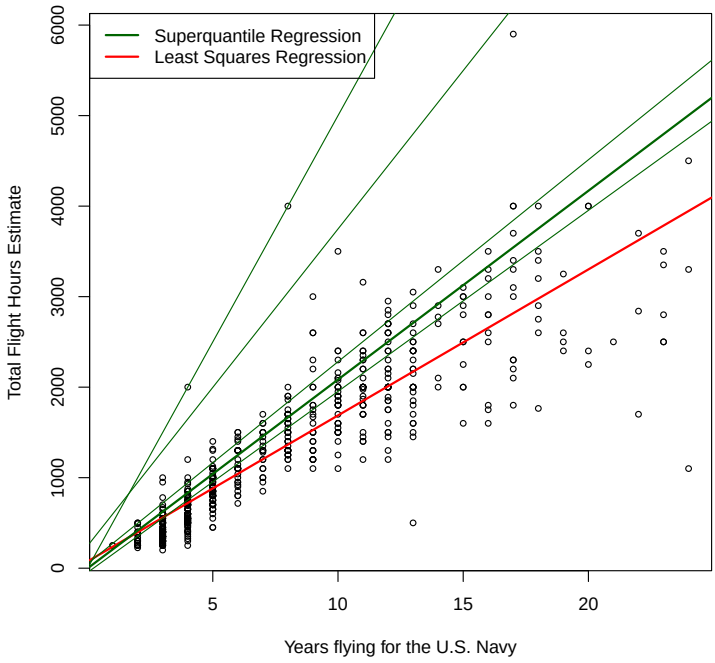
Both least squares and quantile regression functions cross each other at $x_{\text{years}} = 7.91$ years. The observation $(x_{\text{years}}, y) = (2000, 4)$ shifts the least squares regression line upwards for smaller values of x_{years}, while the large number of helicopter pilots with 3 and 4 years flying for the U.S. Navy with low total flight hours shifts the quantile regression model downwards.

In Figure 17(b), we see the least squares regression model and the superquantile regression functions for probability levels $\alpha \in \{0.25, 0.50, 0.75, 0.99, 0.999\}$. We notice that the superquantile regression models for $\alpha = 0.99$ and $\alpha = 0.999$ have higher slopes when compared to the remainder of superquantile regression lines. Even the difference in slopes established by the small increase of 0.009 in α provides us the conclusion that deciding which probability level to use in an analysis is a hard process. Since obtaining these superquantile regression models is not too costly, we consider important to include several choices of probability levels α in any analysis.

From this example we conclude that superquantile regression helps analysts address important questions such as level and trends of the average 1% highest total flight hours (in the case of $\alpha = 0.99$, in Figure 17(b)), understand if deployment rules should be reviewed, and if these cases should be analyzed before reassigning them for



(a) Regression lines for model $f_2(x) = c_0 + c_{\text{years}} x_{\text{years}}$.



(b) Least squares and superquantile regression functions for model $f_2 = c_0 + c_{\text{years}} x_{\text{years}}$, for $\alpha \in \{0.25, 0.50, 0.75, 0.99, 0.999\}$.

Figure 17. Example E: Superquantile regression applied to the U.S. Navy helicopter pilots data.

future deployments.

F. PORTUGUESE SUBMARINERS EFFORT INDEX

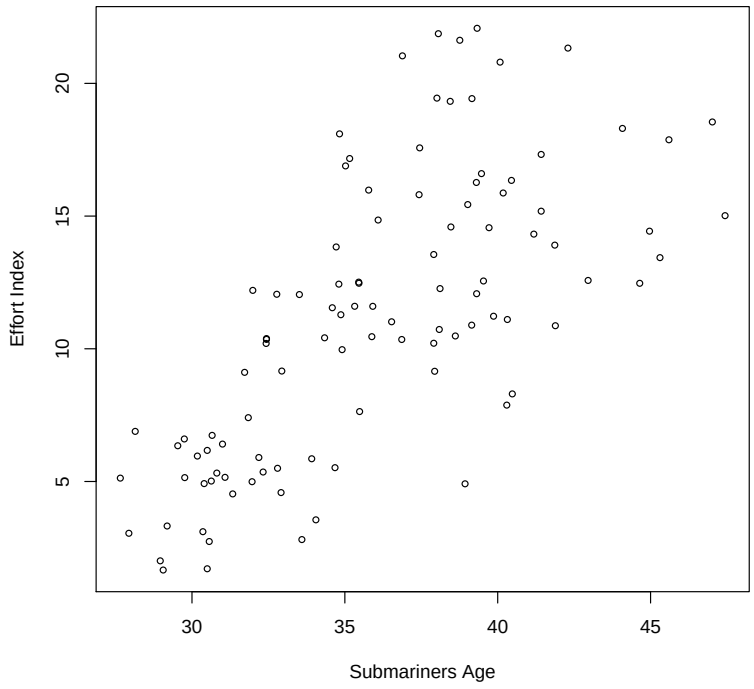
The next example is based on a data set provided by the Portuguese Navy Submarine Squadron. We seek to estimate the random variable Y that represents the effort index of the Portuguese submariners. This index was created as a decision tool to support human resource management inside the Submarine Squadron. Once a sailor becomes a submariner, his career depends mainly on the Submarine Squadron. The Commanding Officer of the Submarine Squadron has the power of assigning a submariner for a mission, if there is the need to embark an extra element or substitute someone onboard. It is crucial to support such decisions with a tool that emphasizes who is more "available" for the mission.

The idea behind this index is to build in the near future a prototype for submariners careers which helps determine selection criteria for future Submarine Squadron personnel recruitment and also understand who has been overemployed.

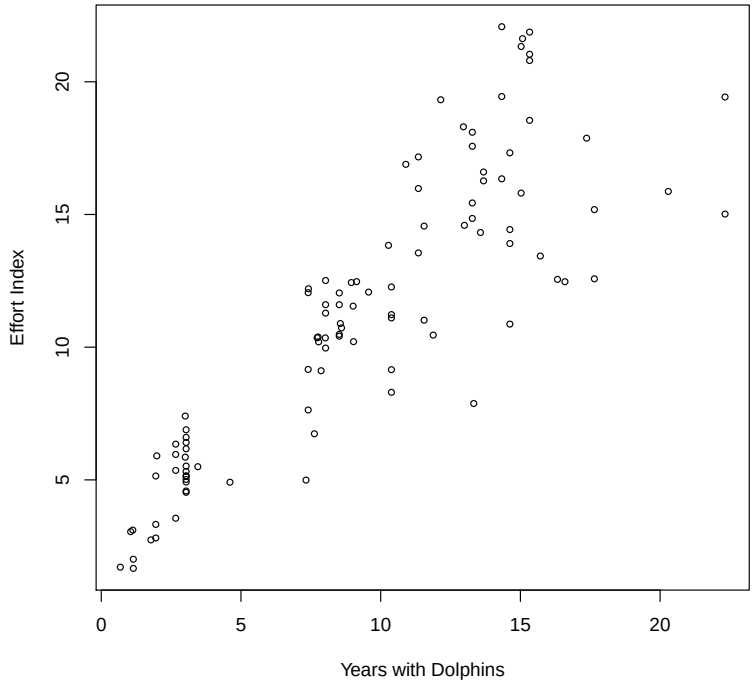
In the data set, we have 103 observations with five possible explanatory variables: years since a submariner has gained the insignia of the Portuguese submarine service (X_{dolphins}), years a submariner has embarked on surface warships (X_{surf}), years a submariner has been ashore (X_{ashore}), total submarine navigation hours (X_{sub}), and submariners age (X_{age}).

Naturally one thinks that age is an important factor that needs to be taken into consideration. The idea of older submariners having more experience due to more training has not always been true, and that issue raised the question of how to quantify training and expertise. Figure 18(a) shows that although age is important, it does not directly translate the effort of a submariner. For example, a 39-year old submariner can have an effort index as low as 5 or as high as 22. Such discrepancies cause discomfort among fellow submariners.

Another factor that is also relevant for designing a prototype for submariners



(a) Effort index versus submariners ages.



(b) Effort index versus years submariners have the dolphins.

Figure 18. Example F: Portuguese submariners effort index against their ages and years they have the submariners insignia.

careers is the number of years a submariner has the insignia of the Portuguese submarine service. Analogously to age, one thinks that the larger this number, the higher the effort index. Figure 18(b) shows how the effort index behaves with the number of years a submariner has the dolphins, and we realize there is an increasing variability among these observations.

For now we consider higher effort indices to be more detrimental than small indices for the completion of the Submarine Squadron mission, i.e., overemploying is considered worse than underemploying a submariner.

One of the goals with this example is to show that superquantile regression helps us better visualize what may cause the discrepancies in effort indices among submariners.

We next observe the possible correlations between variables in the data set. In Figure 19, we have the scatterplot matrix of the data set for some of the explanatory random variables, X_{dolphins}, X_{surf}, and X_{ashore}, against the effort index Y. Here we can observe a linear correlation between the number of years a submariner has the dolphins and the effort index. Since the total submarine navigation hours X_{hours} is a factor considered in the computation of the effort index, their correlation is very high and we do not include this variable in the scatterplot or later in the analysis. We explore several possible models and compare the obtained solution vectors and coefficient of determination results for further analysis.

In Figure 20, we plot the submariners ages against the number of years a submariner has the insignia of the Portuguese submarine service. A small detail that we encounter here is the lack of observations for values of X_{dolphins} between 4 and 7 years. This lack of observations is due to fact that Portugal acquired the Tridente-class submarines in 2010, and the few years prior were dedicated to training the existing submariners to a completely new technology. This process required the Portuguese Navy to delay the submariners course until after the reception of the new assets.

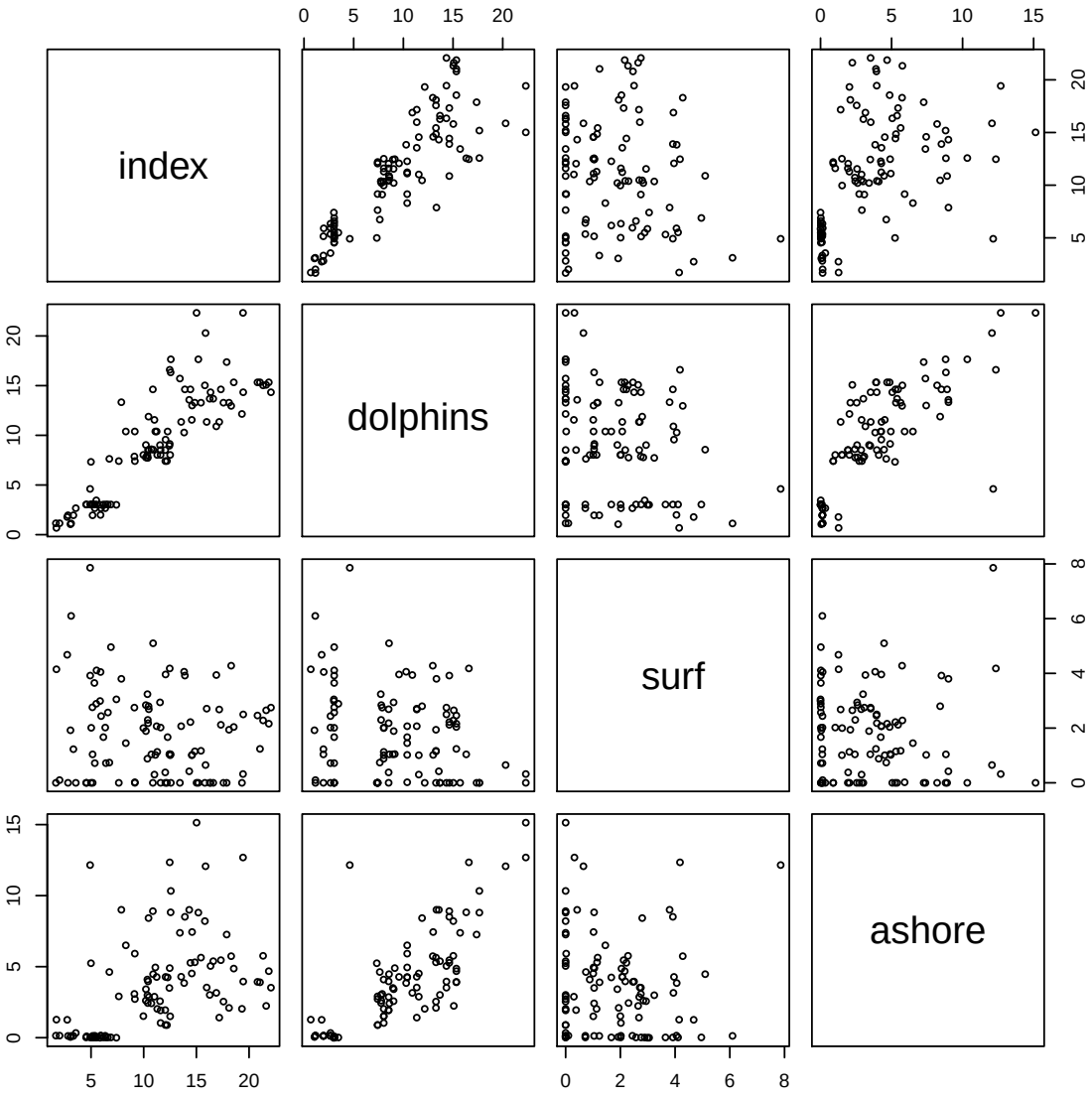


Figure 19. Example F: Portuguese submariners effort index scatterplot matrix.

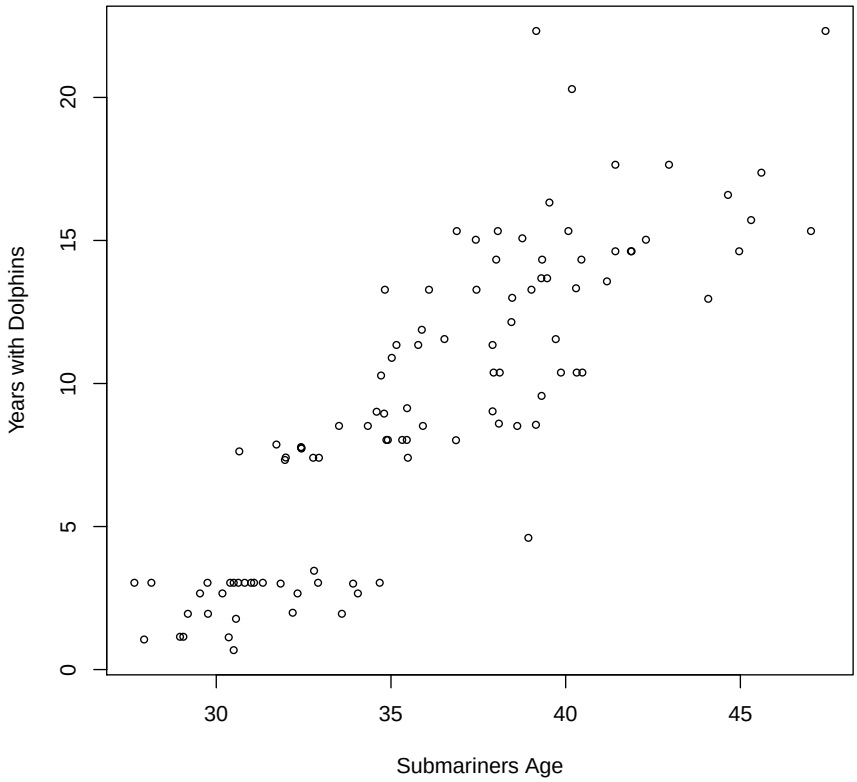


Figure 20. Example F: Submariners ages against the number of years they have the submariners insignia.

We first start with a linear model of the form $f_1(x) = c_0 + c_i x_i$, with $i \in \{\text{dolphins}, \text{surf}, \text{ashore}, \text{age}\}$, i.e., we only include one explanatory variable at a time. We then consider a linear model $f_2 c_{\text{dolphins}} x_{\text{dolphins}} + c_{\text{age}} x_{\text{age}}$. Table 14 presents the obtained solution vectors and the corresponding coefficients of determination, for a probability level $\alpha = 0.75$. The years the submariners embark in surface warships and the number of years they spend ashore between embarks are two explanatory variables that we discard from this point on, because they both play a very negligible role, as determined by $\bar{R}^2_{0.75}$, even though they might be important in conjunction with other explanatory random variables. Rows 6, 11, and 16 of Table 14 report the regression vectors for model f_2. We realize that the coefficient of determination improves in these cases.

Regression	c_0	c_{dolphins}	c_{surf}	c_{ashore}	c_{age}	$\bar{R}^2_{0.75}$
	3.1365	0.8643	—	—	—	0.7452
	11.9111	—	-0.4448	—	—	0.0182
Least Squares	8.3782	—	—	0.7491	—	0.2314
	-17.6369	—	—	—	0.7983	0.4845
	8.1218	0.9918	—	—	-0.1711	0.7512
	2.8690	1.0878	—	—	—	—
	15.8063	—	-0.6190	—	—	—
Quantile	10.7798	—	—	0.7945	—	—
	-19.2554	—	—	—	0.9084	—
	11.7357	1.3037	—	—	-0.3065	—
	2.9811	1.2172	—	—	—	0.5866
	17.6450	—	0.3456	—	—	0.0038
Superquantile	15.4621	—	—	0.5666	—	0.0212
	-27.0234	—	—	—	1.2048	0.2403
	7.3697	1.3430	—	—	-0.1558	0.5939

Table 14. Example F: Regression vectors and $\bar{R}^2_\alpha$ for linear models f_1 and f_2, at a fixed probability level $\alpha = 0.75$.

In Figure 21, we plot the linear model $f_1(x) = c_0 + c_{\text{dolphins}} x_{\text{dolphins}}$ for least squares, 0.60-quantile and 0.60-superquantile regressions. All three obtained regression functions have completely distinct slopes. The blue line representing the quantile regression gives us the notion of where the 40% worst cases are, while the green line representing the superquantile regression model provides us the average of these worst indices.

As stated at the beginning of this example, the orientation of the problem is such that higher effort indices are worse. However and as illustration we believe it is very beneficial to look at the cases where the submariners effort is low and therefore we flip the orientation of this problem for the next figure in order to highlight those cases that should also be taken into consideration. This is a good example where using one of both orientations in solving the problem is possible depending on where the major concern lies. Figure 22 shows the least squares regression model and the 0.75-quantile and 0.75-superquantile regression functions. We add the 0.25-quantile regression fit

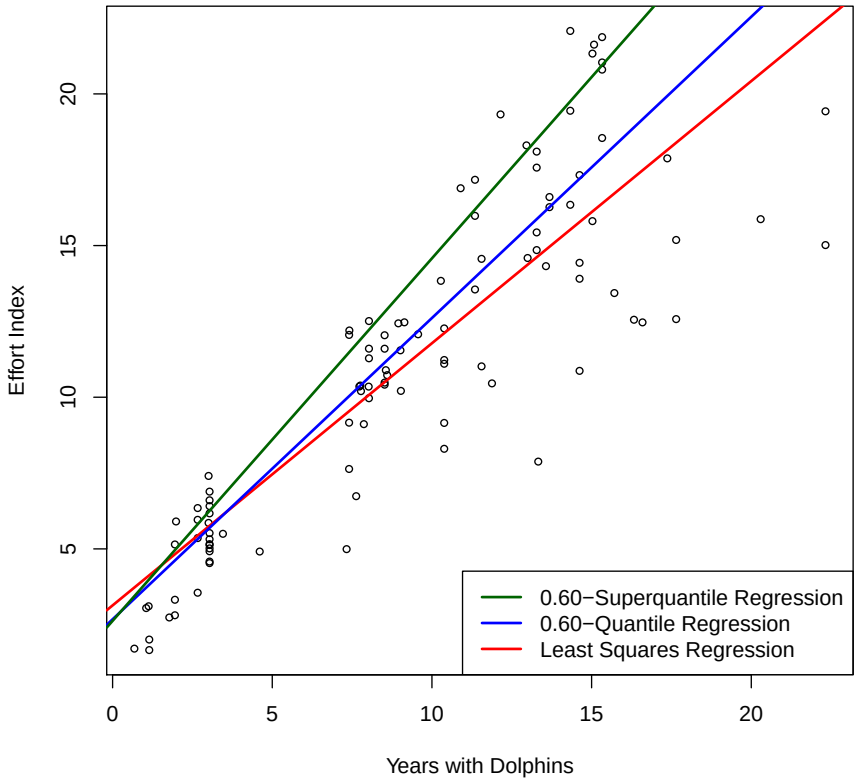


Figure 21. Example F: Regression lines for model $f_1(x) = c_0 + c_{\text{dolphins}} x_{\text{dolphins}}$.

and the new 0.75-superquantile regression function, after flipping the orientation of the problem and solving for the new superquantile regression problem, marked with an asterisk in the legend, and displayed in Figure 22 by a dashed green line. This dashed line has the same meaning as the full green line but for the lowest 25% presented effort indices. We consider that taking care of both ends of the spectrum will expedite the process of smoothing the submariners career, but this is not fully pursued here. We finish using the linear model $f_1(x) = c_0 + c_{\text{dolphins}} x_{\text{dolphins}}$ by showing Figure 23, where clearly the 0.25-superquantile regression model is completely different of the 0.75-superquantile* regression model.

Second, we consider a quadratic model of the form $f_3(x) = c_0 + c_{\text{age}} x_{\text{age}} + c_{\text{age2}} x_{\text{age}}^2$. Table 15 shows the obtained regression vectors and the corresponding co-

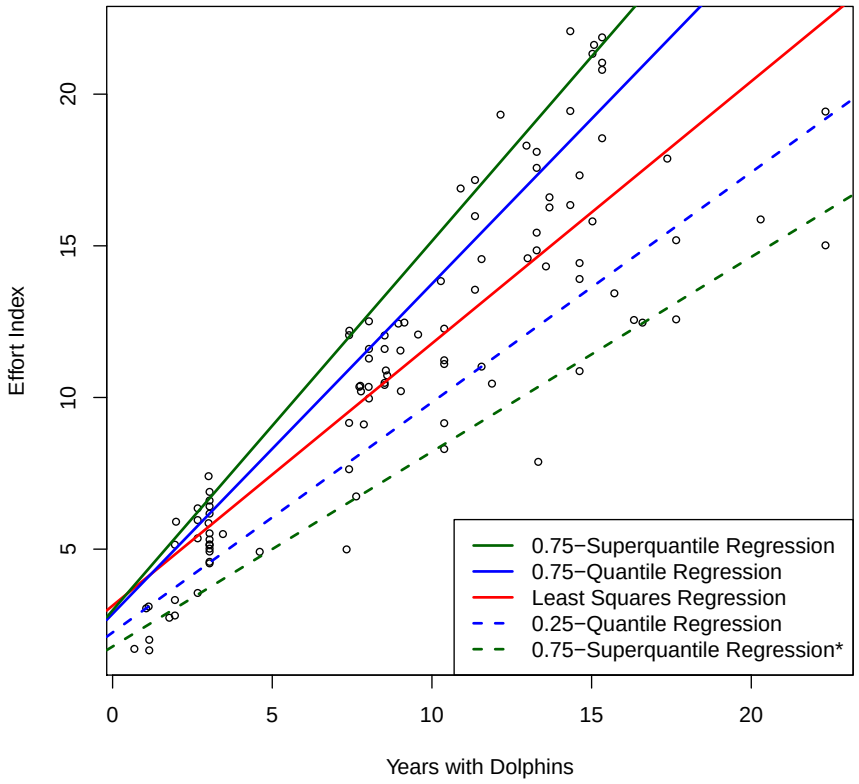


Figure 22. Example F: Least squares, quantile and superquantile regression functions for linear model f_1. An asterisk indicates that the 0.75-superquantile regression function was obtained after reversing the orientation of the original problem.

Regression	c_0	c_{age}	c_{age2}	$\bar{R}^2_{0.75}$
Least Squares	-87.1182	4.6498	-0.05251	0.5442
Quantile	-97.2181	5.2652	-0.0600	—
Superquantile	-126.4859	6.9812	-0.0827	0.3235

Table 15. Example F: Regression vectors and $\bar{R}^2_\alpha$ for quadratic model $f_3(x) = c_0 + c_{\text{age}}x_{\text{age}} + c_{\text{age2}}x^2_{\text{age}}$, with $\alpha = 0.75$.

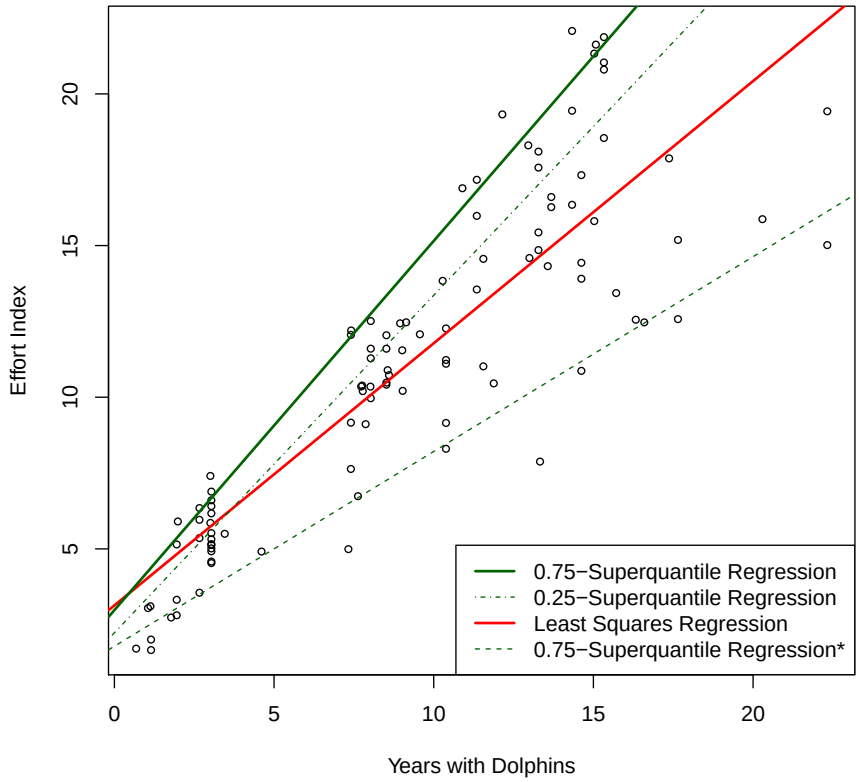
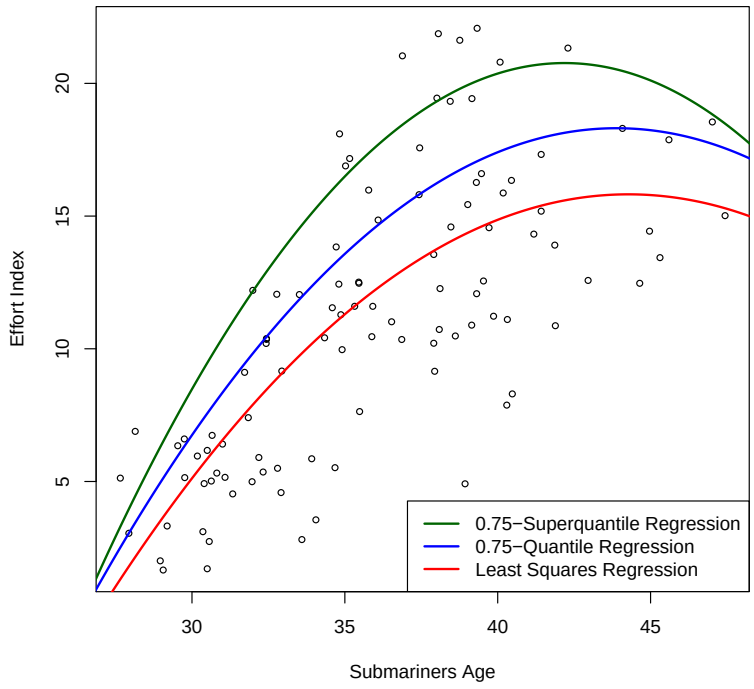
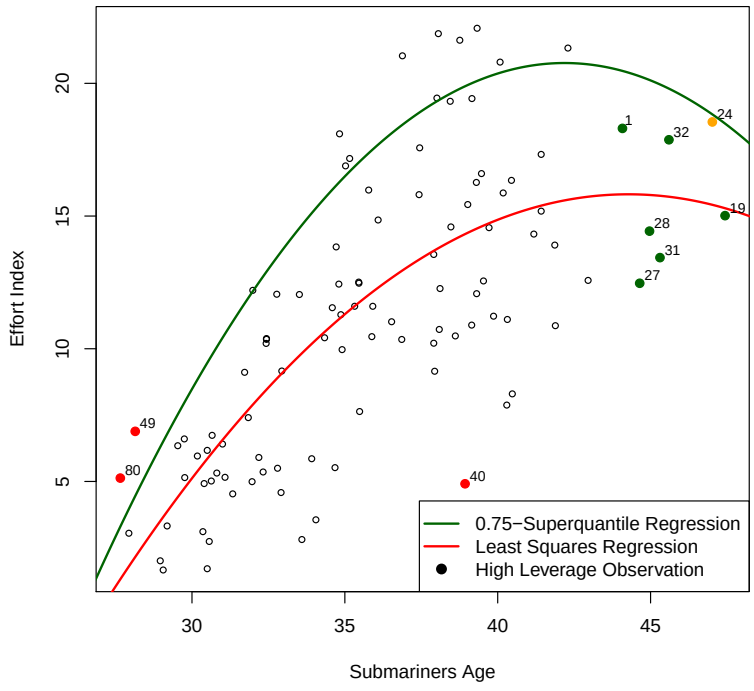


Figure 23. Example F: Different α-superquantile regression functions for linear model f_1. An asterisk indicates that the 0.75-superquantile regression function was obtained after reversing the orientation of the original problem.

efficients of determination, which are larger than those obtained using the linear model. We plot these quadratic models in Figure 24(a), and notice that the 0.75-superquantile regression function captures the effects of the higher effort indices and forms an interesting curvature. To the contrary, the 0.75-quantile regression model does not seem to be affected by such observations and it looks almost parallel to the least squares regression model for a 40-year old submariner. With these comments in mind, we need a different validation analysis tool that helps us understand which observations, if any exist, should be carefully checked for their validity, or should possibly be seen as outliers. Before we finish this example and as seen in Section II.C, we utilize the Cook's distance concept first applied to the case of least squares regression, then to the case of superquantile regression. Since there is more than one possible



(a) Effort index versus submariners ages.



(b) High leverage observations for quadratic model f_3.

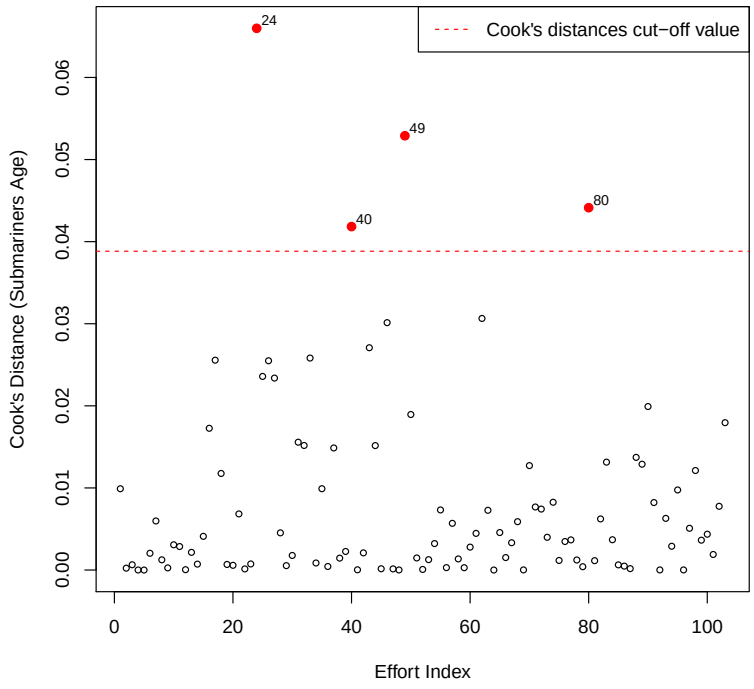
Figure 24. Example F: Quadratic regression models f_3 at probability level $\alpha = 0.75$.

cut-off value for such Cook's distances, we resort to the commonly used formula $4/\nu$ in both cases, least squares and 0.75-superquantile regression, where we have $\nu = 103$ observations in this example. Figure 25(a) shows the Cook's distances for the least squares quadratic model, while Figure 25(b) the Cook's distances for the superquantile regression technique. We clearly see that observations number 24, 40, 49, and 80, emphasized by the red dots in Figure 24(b) and 25(a), are considered high leverage observations for least squares regression. In the context of superquantile regression, we see that observations number 1, 19, 24, 27, 28, 31, and 32, emphasized by the green dots in Figure 24(b) and 25(b), are considered high leverage observations. Curiously, in our example only observation 24 is coincidently considered high leverage for both regression techniques; plotted in orange in Figure 24(b). Another interesting detail consists on where the high leverage observations are located in this same plot. We realize that these observations drive the superquantile regression fit downwards since they influence the 0.75-quantile regression function, and consequently also the resulting 0.75-superquantile regression function as an average of all observations above the quantile regression fit.

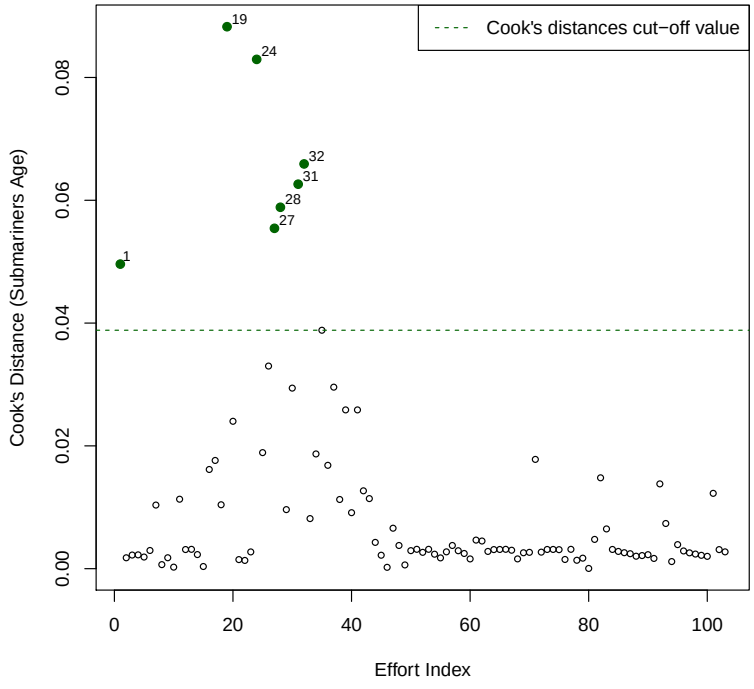
Further analysis should be done for this example but it goes beyond the scope of this dissertation. However we conclude that superquantile regression is an important analysis tool that when used wisely gives the decision maker powerful information on the upper tail of the random variable of concern. With the Cook's distance concept applied to the superquantile regression, we also identify those observations in the data set that influence the obtained fit and that should be checked for validity.

G. UNCERTAINTY QUANTIFICATION

The next example arises in uncertainty quantification of a rectangular cross section of a short structural column, with depth d and width w, under uncertain yield stress and uncertain loads; see Eldred et al. (2011). Assuming an elastic-perfectly plastic material, a limit-state function that quantifies a relationship between loads



(a) Cook's distances for least squares regression.



(b) Cook's distances for superquantile regression.

Figure 25. Example F: Cook's distances for least squares and superquantile regression fits using quadratic model $f_3(x) = c_0 + c_{\text{age}} x_{\text{age}} + c_{\text{age2}} x_{\text{age}}^2$, at $\alpha = 0.75$.

and capacity is described by the random variable

$$Y = -1 + \frac{4X_1}{wd^2X_3} + \frac{X_2^2}{w^2d^2X_3^2}, \tag{IV.1}$$

where the bending moment load X_1 and the axial load X_2 are normally distributed with mean 2,000 and standard deviation 400, and mean 500 and standard deviation 100, respectively, and the material's yield stress X_3, is lognormally distributed with parameters 5 and 0.5, with X_1, X_2, and X_3 independent. We observe that the second term in (IV.1) is the ratio of moment load to the column's moment capacity, and the third term is the square of the ratio of the axial load to the axial capacity. The constant -1 is introduced for the sake of a translation such that positive realizations of Y represent "failure" and negative ones correspond to a situation where load effects remain within the capacity of the column. (We note that the orientation of the limit-state function is switched compared to that of Eldred et al. (2011) for consistency with our focus on "losses" instead of "gains.") We set the width $w = 3$, and the depth $d = 12$.

Model	α	c_0	10^2c_1	10^4c_2	10^4c_3	$\bar{R}^2_\alpha$
	0.999	-0.6797	0.0156	7.9000	-9.1100	0.154
	0.99	-0.8084	0.0150	3.8000	-8.2700	0.190
f_1	0.9	-0.8579	0.0107	1.5900	-7.7000	0.260
	0.75	-0.8705	0.0090	1.0800	-7.5900	0.301
	LS	-0.8827	0.0070	0.5921	-7.7180	0.571
	0.999	-1.0457	1.8640	0.0300	—	0.902
	0.99	-1.0450	1.6182	0.0400	—	0.891
f_2	0.9	-1.0308	1.3393	0.0200	—	0.894
	0.75	-1.0261	1.2595	0.0200	—	0.893
	LS	-1.0179	1.1315	0.0056	—	0.979

Table 16. Example G: Approximate regression vectors and coefficients of determination for superquantile regression with varying α and least squares (LS) regression.

We seek to quantify the "uncertainty" in Y by surrogate estimation. Of course, in this case, this is hardly necessary; direct use of (IV.1) suffices. However, in practice, an analytic expression for a limit-state function, as in (IV.1), is rarely available. One

then proceeds with determining a regression function $f : \mathbb{R}^3 \to \mathbb{R}$, based on a sample of input-output realizations, such that $f(X)$, with $X = (X_1, X_2, X_3)$, approximates Y in some sense. To mimic this situation, we consider a sample of size 50000 drawn independently from X, the corresponding realizations of Y according to (IV.1), and two forms of the regression function. The first model is linear and takes the form

$$f_1(x) = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_3$$

and the second one utilizes basis functions $h_1(x) = x_1/x_3$ and $h_2(x) = (x_2/x_3)^2$ and is of the form

$$f_2(x) = c_0 + c_1 x_1/x_3 + c_2 x_2^2/x_3^2.$$

In view of (IV.1), we expect f_1 to be unable to capture interaction effects between variables and its explanatory power may be limited. In contrast, f_2 uses the correct basis functions, but even then $f_2(X)$ may deviate from Y due to the finite sample size used to determine the regression vector. Table 16 confirms this intuition by showing approximate regression vectors for both models over a range of probability levels α as well as for the least squares (LS) regression. The vectors are obtained in less than 15 seconds by solving $P^{\nu,\mu}_{\text{Num}}$, with $\nu = 50000$, $\mu = 1000$, and Simpson's rule. The last column of Table 16 shows $\bar{R}^2_\alpha$, which is low for f_1 and high for f_2 as expected.

In uncertainty quantification and elsewhere, surrogate estimates such as $f_1(X)$ and $f_2(X)$ are important inputs to further analysis and simulation. Table 17 illustrates the quality of these surrogate estimates in this regard by showing various statistics of $f_1(X)$ and $f_2(X)$ as compared to those of Y. Row 2, Columns 3-10 show estimated mean, standard deviation, superquantiles at 0.75, 0.9, 0.99, 0.999, probability of failure, and buffered probability of failure (see (II.5)) of Y, respectively, using a sample size of 10^7 and standard estimators. Coefficients of variation for these estimators are ranging, approximately, from 10^{-5} for the mean to 0.02 for the probability of failure. Rows 3-6 of Table 17 show similar results, using the same sample, for $f_1(X)$, with $\alpha = 0.999$, 0.99, 0.9, and 0.75, respectively. We notice that as the probability

level α increases, $f_1(X)$ becomes increasingly conservative. In fact, for $\alpha = 0.999$, $f_1(X)$ is conservative in all statistics. Superquantile regression with smaller probability level α fails to be conservative for some "upper-tail" statistics. Interestingly, $f_1(X)$ based on α is conservative for all superquantiles up to and including $\bar{q}_\alpha$ in these tests. These observations indicate that in surrogate estimation the probability level α should be selected in accordance with the superquantile statistic of interest. We can then expect to obtain conservative estimates even for relatively poor surrogates. Row 7 of Table 17 gives corresponding results for $f_1(X)$ under the least squares regression fit. While this fit provides an accurate estimate of the mean (see Column 3), the upper-tail behavior is represented in a nonconservative manner.

Rows 8-12 of Table 17 show comparable results to those above, but for the $f_2(X)$ models. As also indicated in Table 16, $f_2(X)$ is a much better surrogate of Y than $f_1(X)$ and essentially all quantities improve in accuracy. For example, $f_2(X)$ based on superquantile regression overestimates the buffered failure probability only moderately with $\alpha = 0.999$, 0.99, and 0.9, and slightly underestimate with $\alpha = 0.75$; see the last column of Table 17. In contrast, least squares regression underestimates the buffered failure probability substantially even for this supposedly "accurate" model. Of course, least squares regression centers on conditional expectations and as basis for estimating tail behavior may hide potentially dangerous risks.

Model	α	μ	σ	$\bar{q}_{0.75}$	$\bar{q}_{0.9}$	$\bar{q}_{0.99}$	$\bar{q}_{0.999}$	$10^3 p$	$10^3 \bar{p}$
Y	NA	-0.8436	0.0996	-0.7113	-0.6211	-0.3501	0.0091	0.3575	1.052
$f_1(X)$	0.999	-0.1259	0.1297	0.0305	0.0856	0.1868	0.2635	158.1838	376.995
	0.99	-0.4575	0.1027	-0.3370	-0.2963	-0.2225	-0.1669	0	0
	0.9	-0.6940	0.0828	-0.6016	-0.5728	-0.5219	-0.4843	0	0
	0.75	-0.7641	0.0777	-0.6795	-0.6544	-0.6106	-0.5786	0	0
	LS	-0.8439	0.0748	-0.7653	-0.7439	-0.7077	-0.6819	0	0
$f_2(X)$	0.999	-0.7611	0.1647	-0.5381	-0.3961	-0.0053	0.44953	3.4410	9.713
	0.99	-0.7979	0.1431	-0.6042	-0.4808	-0.1413	0.25383	1.4909	4.206
	0.9	-0.8263	0.1184	-0.6660	-0.5640	-0.2831	0.04375	0.4702	1.332
	0.75	-0.8337	0.1113	-0.6830	-0.5870	-0.3229	-0.0155	0.3194	0.899
	LS	-0.8451	0.1000	-0.7097	-0.6235	-0.3864	-0.1104	0.1539	0.440

Table 17. Example G: Statistics of $f_1(X)$ and $f_2(X)$ as compared to those of Y. Columns 3-10 show mean, standard deviation, superquantiles at 0.75, 0.9, 0.99, 0.999, probability of failure, and buffered probability of failure, respectively.

H. SUPERQUANTILE TRACKING

To finish Chapter IV, we return to the first example in Section IV.A, and consider a loss random variable

$$Y = X_1 + X_2\epsilon,$$

where ϵ is a standard normal random variable and $X = (X_1, X_2)$ is uniformly distributed on $[-1, 1] \times [0, 1]$, with ϵ, X_1, and X_2 independent. We consider a regression function of the form $f(x) = c_0 + c_1x_1 + c_2x_2$ and set $\alpha = 0.90$.

We examine conditional values of Y given realizations of $X = (X_1, X_2)$, i.e., superquantile tracking. For $x = (x_1, x_2)$, $Y(x) = Y|X = x$ is normally distributed with mean x_1 and variance x_2^2. Consequently, it is straightforward to compute that $\bar{q}_{0.9}(Y(x)) = x_1 + 1.7550x_2$. Table 2 shows vectors that only track $\bar{q}_{0.9}(Y(\cdot))$ approximately, as c_0, c_1, and c_2 deviate from 0, 1, and 1.755, respectively. In fact, there is in general no guarantee that every regression function f will satisfy $f(x) = \bar{q}_\alpha(Y(x))$ for all x, even for large sample sizes. As indicated by Proposition II.5, however, a superquantile of $Y(x)$ can be estimated by approximating a degenerate distribution of (X, Y) at x.

X range:	$[-1, 1] \times [0, 1]$	$[0.45, 0.55]^2$	$[0.495, 0.505]^2$
$c_0 + 0.5c_1 + 0.5c_2$	(1.349, 1.575)	(1.329, 1.475)	(1.330, 1.473)
c_0	(0.029, 0.123)	(-2.414, 1.784)	(-23.715, 18.329)
c_1	(0.971, 1.075)	(-0.229, 3.597)	(-11.063, 25.656)
c_2	(1.523, 1.975)	(-1.686, 5.186)	(-33.916, 35.701)

Table 18. Example H: Approximate 95% confidence intervals when tracking $\bar{q}_{0.9}(Y(\cdot))$ near $x = (0.5, 0.5)$ using shrinking sampling ranges for X. The correct value $\bar{q}_{0.9}(Y((0.5, 0.5))) = 1.378$.

Table 18 shows such "local" estimates of $\bar{q}_{0.9}(Y(x))$ near $x = (0.5, 0.5)$. Specifically, using $\nu = 500$ we compute c_0, c_1, and c_2 by solving P^ν_{LP} as above, with X sampled uniformly from $[-1, 1]\times[0, 1]$. We repeat these calculations 10 times with independent samples and obtain the aggregated statistics of Column 2 of Table 18. The second row gives an approximate 95% confidence interval for the mean value of $c_0 + 0.5c_1 + 0.5c_2$

across the 10 meta-replications. The interval contains $\bar{q}_{0.9}(Y((0.5, 0.5))) = 1.3775$, but is somewhat wide. Proposition II.5 indicates that sampling from a smaller set $[0.45, 0.55] \times [0.45, 0.55]$ will tend to improve the estimate of $\bar{q}_{0.9}(Y((0.5, 0.5)))$. Column 3 of Table 18 illustrates this effect, by showing results comparable to those of Column 2 and Row 2, but for the smaller interval. As expected, the confidence interval for $c_0 + 0.5c_1 + 0.5c_2$ narrows around the correct value. The last column shows similar results, but now for sampling of X uniformly on $[0.495, 0.505] \times [0.495, 0.505]$. The estimate of $\bar{q}_{0.9}(Y((0.5, 0.5)))$ improves only marginally, with the residual uncertainty being due to the inherent variability in the (relatively small) samples. The narrow sampling interval causes the last estimate to be similar to that obtained by the standard empirical estimate from 500 realizations of $Y((0.5, 0.5))$, which yields the confidence interval $(1.312, 1.462)$.

While sampling on smaller sets gives better local estimates of $\bar{q}_{0.9}(Y(x))$, the global picture deteriorates. The last three rows of Table 18 show corresponding approximate 95% confidence intervals for c_0, c_1, and c_2, respectively. While $c_0 + c_1x_1 + c_2x_2$ generated by the set $[-1, 1] \times [0, 1]$ provides a reasonably good global picture of $\bar{q}_{0.9}(Y(x))$, the smaller sets lose that quality as seen from the wide confidence intervals. In view of the above results, we see that an analyst that can choose "design points," i.e., points x at which to sample $Y(x)$, should balance the need for accurate local estimates with that of global estimates. In fact, even if the primary focus is on estimating $\bar{q}_\alpha(Y(x))$ for a given x, as we see in this example, it may be equally effective to spread the samples of X near x instead of exactly at x, and then obtain some global information about $\bar{q}_\alpha(Y(\cdot))$ too. Our methodology provides a flexible framework for estimating $\bar{q}_\alpha(Y(x))$ even if there is only a small number of realization of $Y(x)$, or even none, available. The estimates are based on realizations of $Y(x')$ for x' near x.

In the next chapter we discuss the conclusions taken from our research and suggest possible future work.

THIS PAGE INTENTIONALLY LEFT BLANK

V. SUMMARY, CONCLUSIONS, AND FUTURE WORK

A. SUMMARY AND CONCLUSIONS

In this dissertation, we develop a novel regression framework, *superquantile regression*, that naturally extends least squares and quantile regressions to contexts where the decision maker is risk averse and is simultaneously concerned about the magnitude of the obtained regression errors. As opposed to squaring these errors or by looking at their signs, this framework for superquantile regression weights larger errors increasingly heavily in a way consistent with a coherent and averse risk measure, the superquantile risk measure. We use superquantiles directly in the regression model and go beyond other generalized regression techniques that approximate conditional superquantiles by various combinations of conditional quantiles, with the only required assumption that the involved random variables have finite second moment.

We utilize the "Fundamental Risk Quadrangle" concept and the connections established therein between distinct measures of a random variable whose orientation is such that upper-tail realizations are unfortunate and low realizations are favorable. We rely on the superquantile-based risk quadrangle and the corresponding relations between measures of deviation, risk, and error applied to the superquantile as the statistic to obtain superquantile regression functions as optimal solutions of an error minimization problem.

Then we develop the fundamental theory for superquantile regression by defining its regression problem as an error minimization problem. We examine existence and uniqueness of the obtained regression functions, and we establish a guaranteed unique regression vector in the cases where the loss random variable and the chosen basis functions are normally distributed with a positive definite variance-covariance matrix. Next we analyze consistency and stability of the regression functions under perturbations due to possible measurement errors and approximating empirical distri-

butions generated by samples of the underlying data. We formulate a deviation-based superquantile regression problem as an equivalent minimization problem of a corresponding measure of deviation taken from the superquantile-based risk quadrangle. This new minimization problem implies computational advantages since it reduces the number of variables and includes a simpler objective function. We also provide rate of convergence results under mild assumptions that allow us to use an approximate superquantile regression problem, based on a sample of the true distribution.

Since any regression framework must be associated with means of assessing the goodness of fit of a computed regression vector, we define three validation analysis tools for quantile and superquantile regressions: the coefficient of determination, the adjusted coefficient of determination, and Cook's distance. We first analyze the formulas for these three validation analysis tools when applied to least squares regression, and translate them into measures of error and deviation in the sense of the mean-based quadrangle. We conclude that these three definitions can be formulated for any generalized regression consisting of minimizing an error random variable.

Concerning computational methods for solving superquantile regression problems, we develop two distinct classes: the primal methods where one solves the superquantile regression problem by means of analytical and numerical integration techniques, and the dual methods where one utilizes the dualization of risk as part of the objective function of the new regression problem that we apply to discrete cases.

In terms of complexity, our results indicate that the dual methods outperform the primal methods in most of the cases, especially for large sample sizes. We compare computational methods by presenting their runtimes and realize that using dual methods is a quite fast process and in fact, for reasonable sample sizes, is not much slower than least squares regression. While the primal method with analytical integration retrieves the exact solutions, it takes too long to run and requires too much memory for sample sizes larger than 1000 observations.

Our results show that superquantile regression is computationally tractable,

offers new insight about the upper-tail-behavior for quantities of interest, and provides a complementary tool for risk-averse decision makers.

B. FUTURE WORK

Similarly to what is done for quantile regression, future work could extend statistical inference and predictive analysis applied to superquantile regression. Also one could further research model validation analysis tools, and address significance testing for superquantile regression. Much research also remains to be done on superquantile tracking. Furthermore, one could build on an R-package to implement superquantile regression and the respective supporting documentation.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

American Optimal Decisions, Inc. (2011). *Portfolio safeguard (PSG) in Windows shell environment: Basic principles*. AORDA, Gainesville, FL.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203–227.

Bertsekas, D. P. (2009). *Convex optimization theory*. Belmont, MA: Athena Scientific.

Billingsley, P. (1995). *Probability and measure* (3rd ed.). New York, NY: John Wiley & Sons.

Borges, C. F. (2011). Discretization vs. rounding error in Euler's method. *The College Mathematics Journal*, *42*(5), 396–399.

Brownlee, K. A. (1960, 2nd ed. 1965). *Statistical theory and methodology in science and engineering* (Vol. 150, pp. 491–500). New York, NY: Wiley.

Cai, Z. & Wang, X. (2008). Nonparametric estimation of conditional VaR and expected shortfall. *Journal of Econometrics*, *147*(1), 120–130.

Chun, S. Y., Shapiro, A., & Uryasev, S. (2012). Conditional value-at-risk and average value-at-risk: Estimation and asymptotics. *Operations Research*, *60*(4), 739–756.

Connor, G. (1995). The three types of factor models: A comparison of their explanatory power. *Financial Analysts Journal*, *15*, 42–46.

Delbaen, F. (2002). Coherent risk measures on general probability spaces. In *Advances in Finance and stochastics* (pp. 1–37). Springer Berlin Heidelberg.

Eldred, M., Swiler, L., & Tang, G. (2011). Mixed aleatory-epistemic uncertainty quantification with stochastic expansions and optimization-based interval estimation. *Reliability Engineering & System Safety*, *96*(6), 1092–1113.

Gilchrist, W. (2008). Regression revisited. *International Statistical Review*, *76*(3), 401–418.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*(494), 746–762.

Hall, P. & Muller, H. G. (2003). Order-preserving nonparametric regression, with applications to conditional distribution and quantile function estimation. *Journal of the American Statistical Association*, *98*(463), 598–608.

Hothorn, T., Kneib, T., & Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 3–27.

Kalinchenko, K., Veremyev, A., Boginski, V., Jeffcoat, D., & Uryasev, S. (2011). Robust connectivity issues in dynamic sensor networks for area surveillance under uncertainty. *Structural and Multidisciplinary Optimization*, *7*(2), 235–248.

Kato, K. (2012). Weighted Nadaraya-Watson estimation of conditional expected shortfall. *Journal of Financial Econometrics*, *10*(2), 265–291.

Knight, J. L., Satchell, S., & Adcock, C. (2005). *Linear factor models in finance*. Elsevier Butterworth-Heinemann.

Koenker, R. (2005). *Quantile regression* (No. 38). Cambridge University Press.

Koenker, R. & Bassett Jr., G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica: Journal of the Econometric Society*, *50*, 43–61.

Krokhmal, P., Zabarankin, M., & Uryasev, S. (2011). Modeling and optimization of risk. *Surveys in Operations Research and Management Sciences*, *16*(2), 49–66.

Lee, S. H. & Chen, W. (2009). A comparative study of uncertainty propagation methods for black-box-type problems. *Structural and Multidisciplinary Optimization*, *37*(3), 239–253.

Leorato, S., Peracchi, F., & Tanase, A. V. (2012). Asymptotically efficient estimation of the conditional expected shortfall. *Computational Statistics & Data Analysis*, *56*(4), 768–784.

Monteiro, R. D. C. & Adler, I. (1989). Interior path following primal-dual algorithms. Part I: Linear programming. *Mathematical Programming*, *44*(1-3), 27–41.

Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence o(1/k2). *Doklady AN SSSR.*, *269*(3), 543–547.

Peracchi, F. & Tanase, A. (2008). On estimating the conditional expected shortfall. *Applied Stochastic Models in Business and Industry*, *24*(5), 471–493.

Phillips, A. S. (2011). *The scope of back pain in navy helicopter pilots* (master's thesis). Retrived from Calhoun http://hdl.handle.net/10945/5795

R Development Core Team (2008). *R: A Language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rockafellar, R. T. & Royset, J. O. (2010). On buffered failure probability in design and optimization of structures. *Reliability Engineering & System Safety*, *95*(5), 499–510.

Rockafellar, R. T. & Royset, J. O. (2014a). Measures of residual risk with connections to regression, risk tracking, surrogate models, and ambiguity (Preprint). *Naval Postgraduate School, Monterey, CA*.

Rockafellar, R. T. & Royset, J. O. (2014b). Random variables, monotone relations, and convex analysis. *Mathematical Programming B*, *148*(1), 297–331.

Rockafellar, R. T. & Royset, J. O. (2014c). Superquantile/CVaR risk measures: Second-order theory (In review). *Naval Postgraduate School, Monterey, CA*.

Rockafellar, R. T., Royset, J. O., & Miranda, S. I. (2014). Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. *European Journal of Operational Research*, *234*(1), 140–154.

Rockafellar, R. T. & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, *2*, 21–42.

Rockafellar, R. T. & Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, *26*(7), 1443–1471.

Rockafellar, R. T. & Uryasev, S. (2013). The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, *18*(1), 33–53.

Rockafellar, R. T., Uryasev, S., and Zabarankin, M. (2008). Risk tuning with generalized linear regression. *Mathematics of Operations Research*, *33*(3), 712–729.

Rockafellar, R. T. & Wets, R. J.-B. (1998). *Variational analysis* (Vol. 317). Berlin: Springer.

Scaillet, O. (2005). Nonparametric estimation of conditional expected shortfall. *Insurance and Risk Management Journal*, *74*, 639–660.

Shor, N. Z. (1985). *Minimization methods for non-differentiable functions*. Berlin: Springer Series in Computational Mathematics.

Trindade, A., Uryasev, S., Shapiro, A., & Zrazhevsky, G. (2007). Financial prediction with constrained tail risk. *Journal of Banking and Finance*, *31*(11), 3524–3538.

Wang, C.-J. & Uryasev, S. (2007). Efficient execution in the secondary mortgage market: a stochastic optimization model using CVaR constraints. *Journal of Risk*, *10*(1), 41–66.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Fort Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California